RESAI TOOLKIT: A FRAMEWOKR FOR CROSSMODALITY BIAS DETECTION.

24_25J_195

BSc (Hons) degree in Information Technology Specializing in Data Science

Department of Information Technology

Sri Lanka Institute of Information Technology

April 2025

RESAI TOOLKIT: A FRAMEWOKR FOR CROSSMODALITY BIAS DETECTION.

(24-25J-195)

Dissertation submitted in partial fulfillment of the requirements for the Bachelor of BSc (Hons) degree in Information Technology Specializing in Data Science

Department of Information Technology

Sri Lanka Institute of Information Technology

April 2025

DECLARATION

We declare that this is our own work and this dissertation does not incorporate without acknowledgement any material previously submitted for a Degree or Diploma in any other University or institute of higher learning and to the best of our knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgement is made in the text.

Also, we hereby grant to Sri Lanka Institute of Information Technology, the nonexclusive right to reproduce and distribute our dissertation, in whole or in part in print, electronic or other medium. We retain the right to use this content in whole or part in future works (such as articles or books).

Signatures of Group Members:

STDUENT NAME	STUDENT ID	SIGNATURE	DATE
H.M.I.K. Dhanawardhana	IT21183690		04/11/2025
E. M. A. M. Ekanayake	IT21387562		04/11/2025
K. M.S. P. Jayawardena	IT21380914		04/11/2025
Mudalige T.N.	IT21208294		04/11/2025

The above candidates have carried out research for the bachelor's degree Dissertation under my supervision.

Signature of the Supervisor: Date: 04/11/2025

ABSTRACT

As Artificial Intelligence (AI) systems increasingly influence critical domains such as healthcare, employment, and media the risk of encoding and amplifying gender bias has become a pressing concern. Traditional fairness metrics often fall short in detecting nuanced, context-dependent biases, particularly across diverse modalities like text, image, audio, and video. This research introduces the RESAI Toolkit, a unified, interpretable framework designed to quantify and analyze contextual gender bias across these four data modalities.

For the text modality, the Context-Aware Bias Metric (CABM) is proposed, integrating cosine similarity, Pointwise Mutual Information (PMI), and sentence-level embedding shifts to assess bias in transformer-based models like BERT. In the image modality**,** the Unified Bias Metric (UBM) combines object-level features (e.g., size, distance, depth) with scene-level semantics using CLIP embeddings and Places365, yielding explainable bias scores supported by SHAP and regression models. The audio modality metric evaluates gender bias by analyzing raw acoustic features such as pitch, amplitude, signal energy, zero-crossing rate (ZCR), voice activity, and the standard deviations of energy, pitch, and amplitude**,** with symbolic and polynomial regression techniques used to generate interpretable bias scores from feature patterns. In the video modality, a multi-dimensional score captures bias through salience-based framing, feature embedding disparity, and representation imbalance across action classes.

Each metric was tested across curated datasets using advanced statistical validation (e.g., Mann-Whitney U, correlation analysis, Shapiro-Wilk test) and machine learning interpretability techniques (e.g., SHAP, PCA, Random Forest, regression modeling). Results reveal that gender bias is contextually driven and manifests uniquely across modalities. The RESAI Toolkit offers a modular, extensible, and reproducible pipeline for detecting and interpreting multimodal bias paving the way for fairer, more accountable AI systems.

ACKNOWLEDGEMENT

This research would not have been possible without the invaluable support, guidance, and encouragement of many individuals and organizations, to whom we are profoundly grateful. First and foremost, we would like to express our sincere appreciation to our supervisor, Dr. Prasanna S. Haddela, for his unwavering support, expert guidance, and continuous encouragement throughout the course of this project. His insightful feedback and mentorship were instrumental in shaping the direction and quality of our work. We are also deeply thankful to our co-supervisor, Ms. Thisara Shyamalee, for her thoughtful suggestions, constructive input, and constant motivation, which significantly contributed to the refinement and success of this research.

We extend our gratitude to the individuals and institutions who supported the data collection process, particularly those who facilitated access to avian datasets. Their contributions formed the foundation upon which this research was built. Finally, we would like to thank the Sri Lanka Institute of Information Technology (SLIIT) for providing us with the necessary resources, facilities, and academic environment that enabled the successful completion of our study. To everyone who supported us directly or indirectly your contributions have been truly meaningful, and we remain sincerely appreciative of your support throughout this journey.

CONTENTS

LIST OF TABLES.

LIST OF FIGURES

LIST OF ABBRIVIATIONS

Abbreviation	Description
AI	Artificial Intelligence
UBM	Unified Bias Metric
OIS	Object Influence Score
SSB	Scene Similarity Bias
CLIP	Contrastive Language–Image Pretraining
YOLO	You Only Look Once
SAM	Segment Anything Model
PCA	Principal Component Analysis
SHAP	SHapley Additive exPlanations
SMOTE	Synthetic Minority Over-sampling Technique
XGBoost	eXtreme Gradient Boosting
CSV	Comma-Separated Values
GPU	Graphics Processing Unit
RGB	Red Green Blue
KDE	Kernel Density Estimation
PWC	PricewaterhouseCoopers
FPR	False Positive Rate
FNR	False Negative Rate
EER	Equal Error Rate
WER	Word Error Rate
CER	Character Error Rate

MSE	Mean Squared Error
NMSE	Normalized Mean Squared Error
RMSE	Root Mean Squared Error
MAE	Mean Absolute Error
BERT	Bidirectional Encoder Representations from Transformers
CABM	Context-Aware Bias Metric
CI/CD	Continuous Integration / Continuous Deployment
CLS	Classification Token (used in BERT)
CNN	Convolutional Neural Network
KDE	Kernel Density Estimation
LLM	Large Language Model
MAC	Mean Average Cosine Similarity
ML	Machine Learning
MLM	Masked Language Modeling
NLP	Natural Language Processing
PCA	Principal Component Analysis
PMI	Pointwise Mutual Information
RF	Random Forest
RNN	Recurrent Neural Network
SHAP	SHapley Additive exPlanations
UBM	Unified Bias Metric (used in literature comparisons)
WEAT	Word Embedding Association Test
WEFE	Word Embedding Fairness Evaluation

1. INTRODUCTION

1.1. Background & Literature Review

1.1.1. Background

Artificial Intelligence (AI) has become deeply embedded in modern decision-making systems across language, vision, audio, and video modalities. As these systems expand into socially sensitive domains such as healthcare, recruitment, surveillance, and legal decision support concerns about their fairness and accountability have intensified. One of the most pressing ethical challenges is gender bias: the systematic tendency of AI models to favor or disadvantage individuals based on gender. This bias, though modality-specific in expression, often stems from a common source: the contextual cues embedded in training data. These cues whether linguistic, visual, acoustic, or behavioral can lead models to encode and reproduce harmful stereotypes, even when high-level accuracy metrics suggest fairness. The four independent research studies in this project tackle this issue from different angles, focusing on the detection of contextual gender bias in text, image, audio, and video AI systems, respectively. This section explores the evolution of gender bias in each modality and introduces the specific methods developed by each group member to quantify and understand it.

Text Modality: Bias in Contextual Word Embeddings

Natural Language Processing (NLP) has evolved rapidly from rule-based and symbolic models to machine learning-driven systems that embed words into dense vector spaces. Early techniques such as Word2Vec [1] and GloVe [2] represented words using fixed-length vectors based on their co-occurrence statistics in large corpora. While these methods enabled meaningful analogies and improved linguistic tasks, they suffered from staticity each word had a single embedding, regardless of context. This was a major limitation when handling polysemous words or phrases whose meanings shift based on sentence structure.

To overcome this, contextual models like ELMo and BERT [3] emerged, producing dynamic embeddings that adapt to surrounding words. However, these innovations brought new risks. Contextual models were shown to encode gender stereotypes more subtly but pervasively. For instance, BERT might associate "He is a doctor" more confidently with professional competence than "She is a doctor," due to biased patterns in its training data [4]. Coreference resolution systems further exemplify this bias, often resolving gender-neutral pronouns to male-coded professions [5].

To address these challenges, H.M.I.K. Dhanawardhana proposed the Context-Aware Bias Metric (CABM), a novel framework that quantifies contextual gender bias in transformer-based embeddings. CABM integrates cosine similarity, Pointwise Mutual Information (PMI), and contextual embedding shifts across sentences to capture both semantic and statistical dimensions of bias. The metric is applied to occupation-related terms using the WinoBias dataset, enabling a nuanced evaluation of how BERT represents gender in real-world contexts. Techniques such as PCA, Random Forest, and SHAP are used to weight and explain the contributing factors behind the computed bias scores. This approach highlights how even fine-tuned language models can reflect deep-seated gender associations, emphasizing the need for context-sensitive auditing tools.

Image Modality: Visual Context and Scene Semantics in Bias

In computer vision, the advent of deep learning particularly convolutional neural networks (CNNs) has revolutionized tasks like object detection, image classification, and facial recognition. Models such as AlexNet [6], YOLO [7], and multimodal systems like CLIP [8] have enabled machines to interpret complex visual data with remarkable precision. However, as these models are deployed in areas like facial recognition, healthcare imaging, and surveillance, they have exhibited troubling gender disparities especially for intersectional groups. These disparities are often not caused by explicit facial features but by visual context, such as background objects or scene layouts.

Studies have shown that women photographed in sports contexts or men in domestic environments are more likely to be misclassified, due to latent associations learned during

training [9]. This phenomenon is known as contextual visual bias, and it remains difficult to detect using traditional fairness metrics.

To address this, E.M.A.M. Ekanayake developed the Unified Bias Metric (UBM), which measures gender bias in visual datasets by analyzing both object and scene-based influences. The UBM consists of two main components: Object Influence Score (OIS), which quantifies spatial, size, and depth-based object impact on gender predictions; and the Scene Similarity Bias (SSB), which uses CLIP and Places365 embeddings to assess semantic similarity between stereotypically gendered environments. The model also employs SHAP for interpretability, allowing practitioners to understand which contextual features contribute most to a biased classification. The evaluation framework integrates PCA, Ridge regression, and XGBoost to ensure robustness across various image configurations. This research advances fairness auditing by introducing a scalable and explainable metric that bridges object-level analysis with high-level scene semantics [10] [11] [12].

Audio Modality: Quantifying Dataset Bias in Speech Signals

Speech-based AI systems are now integral to applications like voice assistants, forensic analysis, and health monitoring for neurological conditions. These systems rely on training datasets that ideally reflect diverse demographic groups. However, in practice, demographic imbalance and recording variability often introduce significant bias. Traditional metrics like False Positive Rate (FPR), False Negative Rate (FNR), Equal Error Rate (EER), and Minimum Detection Cost Function (minDCF) are used to evaluate fairness [13], but these focus on system performance and not on inherent biases in the data itself.

Recognizing this gap, K.M.S.P. Jayawardena proposed a language-independent metric for detecting gender bias in audio datasets, which focuses on raw acoustic features rather than downstream classification errors. The methodology involves extracting statistical audio features such as pitch, amplitude, zero-crossing rate, and signal energy from large-scale speech datasets like LibriSpeech, Common Voice, TED-LIUM, and AMI [14] [15] [16].

Bias scores are then computed based on the differences in feature distributions across gender groups. Controlled bias injections are used to validate metric sensitivity, ensuring that the system can detect imbalances even in cases where traditional error-based metrics appear neutral. This approach provides an essential diagnostic tool for pre-model auditing, offering insights into whether datasets are inherently skewed toward certain demographics before any training occurs [17].

Video Modality: Multidimensional Visual Bias in Human Activity Recognition

Video-based AI systems, particularly in human activity recognition (HAR), are increasingly used in fitness, surveillance, and sports analytics. However, these systems often encode gender bias not just through subject frequency but also through spatial and motion cues embedded in video frames [18], [19]. Addressing this, T.N. Mudalige proposed a Multi-Component Visual Metric to measure gender bias across five dimensions: Size, Centering, Screen Time, Embedding, and Motion Bias.

The methodology uses YOLOv8 for bounding box extraction [20], MediaPipe for pose tracking [20], and SlowFast for action embeddings [22]. These features are normalized and aggregated into directional, magnitude, and PCA-weighted bias scores. When tested on the HAA500 dataset [23], the metric uncovered strong gender skew in classes like *yoga_dancer* (female-leaning) and *tennis_serve* (male-leaning). Notably, bias was detected even in gender-balanced classes, highlighting how visual framing and motion style influence representational fairness.

This framework offers an interpretable and scalable tool for bias auditing in HAR systems and supports fairer video-based AI development.

Unifying Context: Why Bias is a Multimodal, Contextual Phenomenon

Despite the differences in data type, architecture, and application, a shared insight emerges across these four studies: gender bias is contextually driven. In text, it appears in sentence structure and token embeddings; in images, through object proximity and scene setting;

in audio, via pitch and speaking style; and in video, through spatial framing and action associations. These contextual cues often go unnoticed by traditional bias metrics, which tend to focus on surface-level attribute distributions or outcome parity.

Existing fairness tools such as WEAT [1], REVISE [10], AIF360 [11], and Fairlearn [12] provide important foundational methods for auditing bias, but they typically fall short when applied to multimodal, high-dimensional data. They lack the ability to interpret how context rather than just content shapes model predictions. This limitation is particularly concerning in sensitive applications where biased decisions can reinforce societal inequities.

By designing modality-specific metrics that account for contextual nuances, the individual studies in this research project contribute to a deeper understanding of how bias manifests and how it can be quantified. These contributions not only help identify existing issues but also pave the way for more equitable AI systems across disciplines.

1.1.2. Literature Review

Text-Based Systems (NLP)

The field of Natural Language Processing (NLP) has witnessed significant progress from static word embeddings to contextual language models. Early studies like the Word Embedding Association Test (WEAT) by Caliskan et al. [4] laid the foundation for quantifying semantic biases in static models like Word2Vec [1] and GloVe [2]. These embeddings, while powerful, assigned a single vector per word, failing to capture meaning variation based on context. As contextual models like BERT [3] emerged, new challenges surfaced. Gonen and Goldberg (2019) demonstrated that even after debiasing, residual gender subspaces persist in contextual embeddings, subtly influencing model decisions. Zhao et al. [5] showed that coreference resolution systems disproportionately associate male pronouns with professional roles, revealing deep-seated bias in sentence-level behavior.

Beyond WEAT, newer metrics like Mean Average Cosine Similarity (MAC), Pointwise Mutual Information (PMI), and SAME Score (from StereoSet) attempted to address contextual nuances. However, these tools often lack transparency and struggle to generalize across tasks. The WEFE framework by Badilla et al. aimed to unify evaluation across embedding types, yet it too focused on static representations and lacked sentence-level sensitivity. To bridge this gap, Dhanawardhana proposed the Context-Aware Bias Metric (CABM), which integrates cosine similarity, PMI, and BERT-based embedding shifts with PCA and SHAP to detect nuanced contextual gender bias in occupational terms.

Image-Based Systems (Computer Vision)

Computer vision has faced increasing scrutiny for encoding gender and racial biases. Buolamwini and Gebru [9] exposed demographic disparities in commercial face analysis systems through their Gender Shades study, revealing error rate gaps exceeding 30% between dark-skinned women and light-skinned men. Tools like REVISE [10] assess object-gender co-occurrence in datasets like COCO but lack the spatial and semantic depth required to detect contextual influence.

Meister et al. [18] further investigated the presence of gender artifacts in widely used image datasets, identifying how subtle cues like clothing style, posture, and background elements unintentionally reinforce gender stereotypes, thereby skewing classification outcomes even in state-of-the-art models.

Recent research by Sabir and Padró [19] further illustrates how object-gender correlations such as the presence of lipstick in images can amplify gender bias in classification systems, highlighting how seemingly neutral objects may carry strong social connotations. General-purpose toolkits like AIF360 [11] and Fairlearn [12] offer statistical parity metrics but are optimized for tabular inputs, making them unsuitable for complex visual features.

Interpretability methods like Grad-CAM [20] and Score-CAM [21] visualize attention saliency, showing what parts of an image influenced a model's prediction. However, they offer no quantitative bias score. Ekanayake's Unified Bias Metric (UBM) responds to this shortfall by quantifying how object proximity and scene semantics influence gender classification. The UBM combines an Object Influence Score (OIS), which captures depth and spatial weight, with a Scene Similarity Bias (SSB) using embeddings from CLIP [8] and Places365. SHAP explanations and ridge regression models enable interpretable diagnostics across datasets.

Audio-Based Systems

In audio AI systems, fairness has typically been assessed using performance-based metrics like False Positive Rate (FPR), False Negative Rate (FNR), Equal Error Rate (EER), and Minimum Detection Cost Function (minDCF) [13]. These indicators measure prediction errors across demographics but cannot reveal dataset-level bias prior to training. Metrics like G2mindiff, G2avgRatio, and G2avg log Ratio [17] offer relative performance disparity measurements, comparing how groups perform against a baseline. While useful, they are influenced by downstream model choices.

To isolate inherent bias, Jayawardena proposed a statistical method that focuses on raw audio features such as pitch, signal energy, and zero-crossing rate. This approach allows pre-training bias detection by examining whether specific demographic groups are over- or underrepresented in the acoustic feature space. Validation using datasets like LibriSpeech, TED-LIUM, and Common Voice confirms that traditional metrics may overlook latent biases that originate from recording or demographic imbalance.

Video-Based Systems (Human Activity Recognition)

Video-based AI systems, particularly those used in human activity recognition (HAR), introduce unique challenges for fairness due to the temporal, spatial, and motion dynamics encoded in video data. Traditional image-based bias metrics fall short in this domain as they overlook how pose trajectories, screen presence, and framing contribute to representational skew. Mudalige addresses this gap with a Multi-Component Visual

Metric, a novel framework designed to detect gender bias across five visual dimensions: Size Bias, Centering Bias, Screen Time Bias, Embedding Bias, and Motion Bias.

Metric uses a custom feature extraction pipeline incorporating YOLOv8 for bounding box analysis [20], MediaPipe for pose estimation [21], and the SlowFast model for action embeddings [27]. These features are normalized and combined into directional, magnitude, and PCA-weighted scores to assess both the strength and orientation of gender bias in activity videos. Empirical results on the HAA500 dataset [8] reveal strong gender asymmetries across sports and fitness activities. For instance, *yoga_dancer* and *yoga_cat* skew female, while *tennis_serve* and *badminton_serve* skew male—even in categories with balanced participation. These findings suggest that bias arises not only from frequency but also from visual framing and motion styles. MCVM offers a scalable, interpretable framework for fairness auditing in video datasets, extending gender bias analysis into the temporal domain.

Summary

Across all modalities, current tools fail to account for contextual dependencies—how gender bias emerges not just from subject features but also from background, structure, or co-occurrence. Traditional metrics focus on accuracy parity or static associations, missing deeper semantic patterns. The individual contributions from this project—CABM, UBM, audio statistical bias detection, and video saliency-aware debiasing—address this gap with domain-specific and interpretable metrics. They collectively demonstrate that fairness in AI must evolve toward multimodal, context-aware evaluation frameworks.

1.2.Research Gap

Despite rapid progress in fairness-aware AI research, a significant gap persists across modalities text, image, audio, and video in detecting and quantifying contextual gender bias. Most existing fairness toolkits, such as WEAT, REVISE, AIF360, and Fairlearn, focus predominantly on static features, demographic distributions, or performance-based

error metrics, which overlook the nuanced, often subtle manifestations of bias driven by contextual factors like spatial layout, co-occurrence patterns, or syntactic framing.

In Natural Language Processing (NLP), traditional bias detection methods such as WEAT and MAC are limited to static word embeddings and fail to address how bias dynamically emerges in contextual embeddings like BERT. Sentence-level behavior, polysemy, and token interaction are largely ignored, undermining the interpretability and generalizability of results. Moreover, frequency-sensitive metrics like PMI can misrepresent associations in low-resource settings, leading to instability. Current tools also lack transparency, offering raw bias scores without explanations of the contributing factors, as critiqued by Badilla et al. (2020). Intersectional biases further complicate detection, necessitating multi-dimensional and explainable tools that can operate across transformer layers and real-world corpora.

In image-based systems, the majority of bias evaluations focus on subject-centered facial features or demographic imbalance, overlooking the influence of visual context such as object salience, scene layout, and object-scene interaction. While tools like REVISE offer demographic and co-occurrence analysis, they fail to quantify bias caused by latent spatial and semantic cues. Studies by Sabir and Padró (2023) and Meister et al. (2023) show that contextual artifacts lipstick, background, posture serve as unintended gender signals, but their methods lack scalable, numerical metrics. Furthermore, while interpretability tools like Grad-CAM and Score-CAM visualize localized saliency, they are not designed to generalize across datasets or quantify systemic contextual bias.

In audio-based systems, bias detection remains largely performance-driven, emphasizing metrics like EER or FNR. These methods ignore dataset-level issues such as class imbalance, demographic underrepresentation, and signal feature disparities. Most fail to isolate bias originating from acoustic features before model training begins. As highlighted by Hutiri et al. (2024), methodological flaws in bias evaluation for speaker verification systems persist. There is a clear need for a statistical approach to detect

imbalances in raw audio features like pitch, energy, and ZCR features directly influenced by demographic factors yet overlooked by traditional metrics.

Video-based action recognition systems present a compounded challenge due to the interplay of spatial and temporal features. Existing methods often rely on clip count parity or post hoc performance gaps, ignoring the subtle ways gender bias infiltrates motion patterns, framing, and salience. Deep embeddings, which encode both spatial appearance and movement, often carry gendered associations that skew recognition outputs. Studies like Wang et al. (2019) and Meister et al. (2023) identify this bias but lack tools to systematically measure it across datasets. The underutilization of causal graphs and embedding-level salience further limits bias interpretability. Additionally, there is a lack of unified frameworks that integrate multiple bias signals such as representation imbalance, attention salience, and feature separation into a composite metric capable of correlating dataset fairness with model behavior.

Across all domains, most existing solutions treat bias as a surface-level issue and fail to capture the full pipeline from data structure to decision-making logic. There is a pressing need for explainable, domain-specific, and context-aware fairness metrics that reflect the actual complexity of gender bias in AI. The individual studies presented in this project contribute to bridging these gaps by proposing targeted, interpretable tools tailored to the unique challenges of each modality, paving the way for a holistic understanding of bias in multimodal systems.

1.3.Research Problem

The proliferation of AI across text, image, audio, and video domains has triggered an urgent call for rigorous fairness evaluation. However, existing bias detection strategies remain fragmented and modality-specific, failing to provide a unified or interpretable framework for identifying contextual gender bias. Across all modalities, the prevailing challenge lies in the detection and quantification of gender bias induced not only by the subjects themselves but also by the contextual environments—linguistic structures, scene elements, acoustic properties, and spatiotemporal framing. Each modality exhibits unique

pathways through which bias is encoded, yet traditional metrics fail to capture this diversity.

In the text domain, most bias evaluations focus narrowly on static word associations, overlooking how contextual word embeddings like those generated by BERT shift meaning depending on sentence structure. Metrics such as WEAT, MAC, and PMI provide limited insight into sentence-level embedding variations. Consequently, there is a critical need for a metric that integrates semantic similarity, statistical co-occurrence, and contextual embedding dynamics to assess bias holistically and explainably at the sentence leel.

For image-based systems, current fairness tools primarily measure label co-occurrence or demographic representation without accounting for contextual visual features. Studies have shown that background objects, object depth, and spatial relationships significantly influence gender predictions, yet no tool systematically quantifies their impact. The lack of a unified, interpretable metric to capture how object salience and scene semantics reinforce stereotypes poses a major barrier to equitable computer vision systems.

Similarly, in audio systems, performance-based fairness metrics such as EER and minDCF obscure inherent dataset-level gender imbalances. There is an absence of tools capable of isolating gender bias in acoustic feature distributions such as pitch, zero-crossing rate, or signal energy before model training. The need is clear: a statistical diagnostic tool that functions independently of downstream classification tasks and reveals demographic skews embedded in voice data.

In video-based human action recognition, biases often arise from framing, scene composition, and motion encoding. These elements introduce visual salience disparities and embed gender associations within feature-level representations. Yet, few studies provide metrics that can evaluate how such contextual bias affects model attention and classification boundaries. Furthermore, existing dataset audits do not correlate detected biases with downstream performance metrics like accuracy or error rates, limiting their practical relevance.

Therefore, the central research problem addressed in this work is: How can we develop robust, interpretable, and context-aware multimodal metrics that detect gender bias across text, image, audio, and video AI systems by integrating semantic, spatial, acoustic, and temporal context into unified frameworks? Addressing this question requires the design of domain-specific yet generalizable bias metrics that move beyond simplistic parity checks to capture the nuanced and often latent ways in which context shapes biased decision-making. The intended outcome is a set of tools that not only identify bias but also provide actionable insights for mitigation, thereby enabling the development of fairer AI systems across modalities.

1.4.Research Objectives

To address the complex and multifaceted issue of gender bias across diverse AI modalities text, image, audio, and video this research proposes a unified objective framework tailored to the strengths and limitations of each data type. The overarching goal is to develop, implement, and validate modality-specific yet conceptually aligned metrics for detecting contextual gender bias in data used to train AI systems. These metrics emphasize explainability, statistical rigor, and practical relevance, facilitating better fairness audits across disciplines.

The central objective of this study is to introduce a suite of contextual bias metrics that assess how hidden cues linguistic, visual, acoustic, or spatiotemporal contribute to gender misrepresentation in AI model training datasets. By focusing on contextual rather than solely demographic indicators, the research advances current methodologies and delivers tools that detect subtler, more insidious forms of gender bias.

Specifically, for Natural Language Processing (NLP), the objective is to create the Context-Aware Bias Metric (CABM), which integrates semantic similarity (via cosine similarity), statistical co-occurrence (via Pointwise Mutual Information), and sentence-level contextual shifts (via BERT embeddings). This will be supported by dimensionality reduction (PCA), feature attribution (SHAP), and statistical testing (Mann-Whitney U, Chi-Square).

For image datasets, the Unified Bias Metric (UBM) is developed to combine object-level influence captured through Object Influence Score (OIS) and scene-level semantics captured through Scene Similarity Bias (SSB) using CLIP and Places365 embeddings. Machine learning models like Ridge Regression and XGBoost enhance score accuracy, while SHAP plots and comparative visualization aid interpretability.

In audio analysis, the objective is to design a language-independent metric that quantifies dataset-level gender bias based on raw acoustic features like pitch, amplitude, and signal energy. It further investigates how speaker attributes like race and language interact with these features and aims to establish robust criteria for unbiased data representation.

For video-based systems, the objective is to develop a visual bias metric that detects gender bias in human activity recognition datasets using five features: size, centering, screen time, embedding, and motion. These are extracted through models like YOLOv8, MediaPipe, and SlowFast, and combined using standardized scores and aggregate measures. The aim is to reveal bias in visual framing and movement patterns, even in datasets with balanced gender counts, offering a practical tool for fairness evaluation in spatiotemporal data.

Together, these objectives lay the groundwork for developing robust, interpretable, and context-sensitive tools that move AI systems toward greater fairness and equity. Each modality-specific metric not only supports pre-model dataset audits but also contributes to longitudinal fairness tracking across the AI lifecycle.

2. METHODOLOGY

To effectively address gender bias in artificial intelligence systems, it is essential to evaluate the bias present at the data level across diverse modalities. This methodology section presents a comprehensive set of bias detection metrics specifically designed for four key modalities: textual embedding, image, audio, and video. Each metric is developed to capture modality-specific factors contributing to gender bias, while aligning under a shared goal of detecting and quantifying inherent bias in datasets, independent of downstream model performance. This is done under four components:

1. Bias in Word Embeddings proposes a metric for measuring gender bias in contextualized word embeddings (e.g., BERT), focusing on how gendered words interact within different sentence contexts.
2. Bias in Image Datasets addresses gender bias using contextual visual elements like objects, scenes, and spatial relationships that co-occur with gendered representations in images.
3. Audio Bias Score introduces a novel metric for detecting gender bias in audio datasets by analyzing raw audio features such as pitch, energy, and speech activity across gender groups, without relying on model predictions.
4. Bias in Human Action Recognition explores gender representation in action recognition datasets, using multi-dimensional analysis of action types, frequencies, and gender distributions to detect skewness and stereotype propagation.

Together, these four methodologies lay the foundation for a multimodal bias detection framework, enabling holistic assessment of gender bias at the dataset level. Each section below will detail the metric construction, dataset selection, feature extraction, and evaluation strategies specific to its modality.

2.1. Developing A Metric to Detect Gender Bias in Contextual Word Embeddings.

2.1.1. . Introduction to Methodology

This chapter details the methodological framework developed to effectively measure and analyze gender bias within contextual word embeddings. Building on the identified research gaps and objectives outlined in the Introduction chapter, this methodology systematically integrates multiple computational dimensions into a unified metric, thereby addressing the limitations present in existing bias detection approaches.

The primary goal of this research is to develop a novel, comprehensive, and interpretable metric Context-Aware Gender Bias Score (CABM) capable of accurately detecting gender bias in contextual embeddings, specifically those generated by transformer-based models such as BERT. This metric is designed to overcome significant drawbacks found in traditional bias measurement techniques, particularly their inability to capture nuanced, sentence-level contextual biases effectively. To achieve this overarching objective, the research specifically pursues the following detailed sub-objectives:

The methodology begins by extracting these critical bias-related features from the widely recognized WinoBias dataset, which includes carefully structured sentences designed to test both stereotypical and anti-stereotypical associations in semantic and syntactic contexts. Following extraction, the research employs three advanced statistical and machine-learning-based weighting approaches Principal Component Analysis (PCA), PCA combined with Random Forest Feature Importance, and Random Forest combined with SHapley Additive exPlanations (SHAP) to optimally integrate these features into the CABM. PCA provides interpretability through dimensionality reduction, PCA combined with Random Forest captures non-linear interactions among features, and SHAP combined with Random Forest provides clear interpretability by quantifying individual feature contributions explicitly.

The resulting CABM, computed using weights derived from these methodologies, is

rigorously validated through statistical tests, including the Mann-Whitney U test, Shapiro-Wilk test, and Chi-Square test, ensuring its robustness and generalizability. Comparative analyses against traditional metrics like WEAT, MAC, and PMI further demonstrate the enhanced effectiveness of CABM. Additionally, visualization methods such as heatmaps and distribution plots are developed to clearly communicate bias patterns, making the results accessible to both technical experts and general stakeholders.

Finally, the methodology ensures reproducibility and extensibility, facilitating future adaptation to other types of social identity biases (e.g., race or nationality) and alternative transformer-based language models. This comprehensive approach not only advances bias detection accuracy but also significantly improves interpretability and applicability in real-world NLP applications.

2.1.2. System Architecture

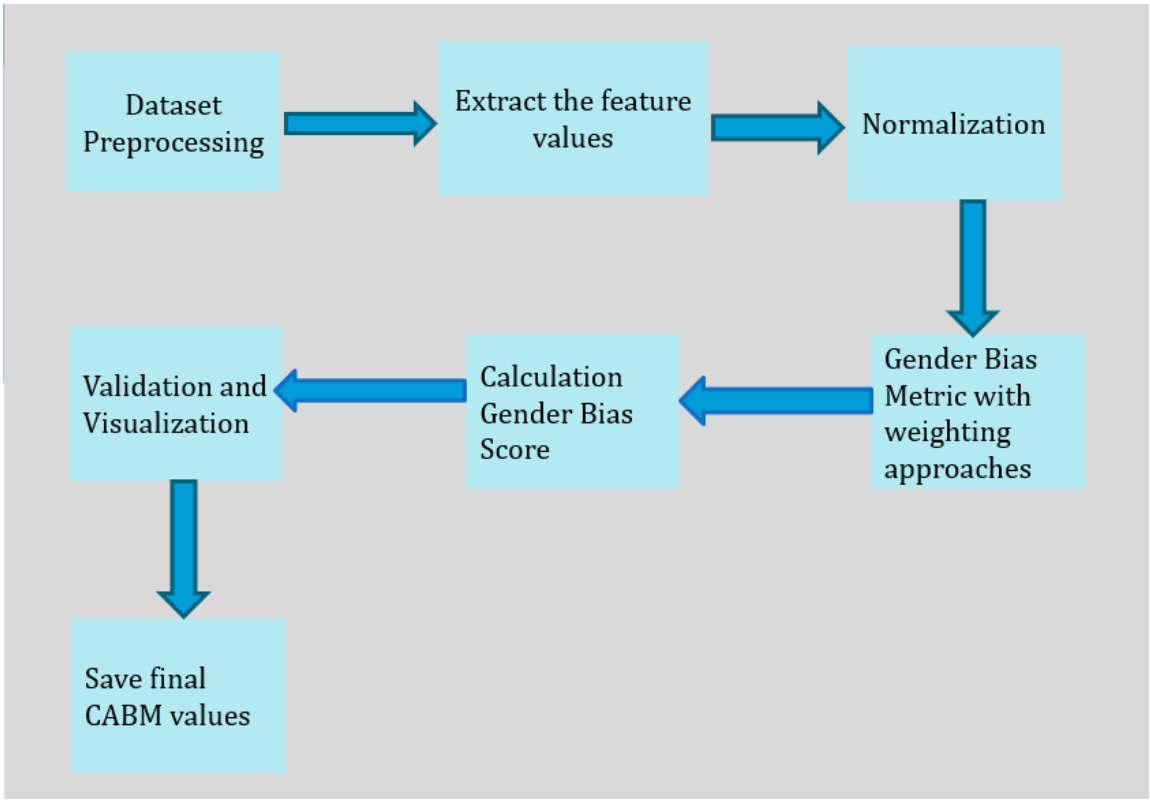

Figure 1: Calculating CABM System Architecture

2.1.3. Data Collection and Preprocessing

This section describes the data sources and preprocessing methods used to ensure robust and accurate bias analysis using the Context-Aware Bias Metric (CABM). It includes a detailed explanation of the dataset selected (**WinoBias**), its structure and significance, and the preprocessing steps necessary to prepare data for embedding extraction and subsequent feature computations.the collection, labeling, and preparation of the datasets.

Dataset: WinoBias

The WinoBias dataset, introduced by Zhao et al. [22], was specifically selected for this research due to its effectiveness in evaluating gender bias within coreference resolution tasks. Designed to identify and measure biases in NLP models, WinoBias provides sentence pairs carefully structured around stereotypical and anti-stereotypical gender roles, creating ideal scenarios to test and quantify bias in contextual embedding models such as BERT.

The dataset comprises two types of sentence pairs:

- Type 1 (Semantic Bias): These sentence pairs are crafted with clear semantic cues that may align or conflict with common gender stereotypes, helping assess the extent to which contextual embeddings encode semantic stereotypes.

Example (Semantic Type 1):

  - Pro-stereotypical: "The doctor told the nurse that he was late."
  - Anti-stereotypical: "The doctor told the nurse that she was late."

- Type 2 (Syntactic Bias): These sentence pairs utilize syntactic structures designed to test if biases persist even when syntactic clues dominate semantic ones, examining the subtle influence of syntactic structure on model behavior.

Example (Syntactic Type 2):

- Pro-stereotypical: "The nurse notified the doctor that her shift was ending."
- Anti-stereotypical: "The nurse notified the doctor that his shift was ending."

Each sentence explicitly targets gender biases associated with occupations, roles, and pronoun references, providing a robust basis for measuring both overt and subtle gender biases in language embeddings.

Final Dataset Composition

After preprocessing, filtering, and gender annotation, the finalized dataset contained 927 contextual sentences involving 39 unique occupations, each embedded within either male- or female-gendered sentence frames. The gender distribution was balanced, with 480 male and 447 female examples, ensuring fair representation across gendered contexts. Each sentence was further annotated with its corresponding occupation, pronoun, and a masked sentence version for embedding extraction. This curated dataset was then passed through the CABM pipeline for feature extraction, bias computation, and weighting analysis across all three metric variants.

Sentence Tokenization

The preprocessing pipeline begins with tokenizing sentences into word-level units. Sentence tokenization involves breaking down each sentence in the dataset into smaller linguistic components (tokens), allowing for subsequent embedding extraction and analysis. For consistency and reproducibility, the tokenization process uses the BERT tokenizer (WordPiece tokenizer) provided by the Hugging Face Transformers library, which ensures alignment with BERT's embedding vocabulary and model requirements.

Example:

Original Sentence:

- "The doctor told the nurse that he was late."

Tokenized Version (BERT Tokenizer):

- ["[CLS]", "The", "doctor", "told", "the", "nurse", "that", "he", "was", "late", ".", "[SEP]"]

Including special tokens "[CLS]" and "[SEP]" helps BERT understand sentence boundaries and facilitates accurate embedding extraction.

```
{'input_ids': tensor([[  101,  1996,  5553, 15660, 16360, 20026,  5685,  2098,  1996, 17907,
          2138,  2002,  2081,  1037,  6707, 15242, 17397,  1012,   102]]),
 'token_type_ids': tensor([[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]]),
'attention_mask': tensor([[1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1]])}
```

Figure 2:tokenized dataset

Getting Embeddings using BERT

After tokenization, the sentences were processed using the pre-trained BERT model (**bert-base-uncased**), leveraging its deep bidirectional Transformer architecture [23]. BERT was chosen because of its superior ability to capture contextual semantics dynamically, providing accurate representations of words based on their specific sentence contexts.

To generate embeddings:

- Each tokenized sentence is input into BERT.
- The BERT model outputs a 768-dimensional embedding for each token.
- Embeddings corresponding to occupation-related words ("doctor," "nurse," "engineer," etc.) are specifically extracted, as they are central to bias analysis.

Example embedding extraction:

- Input sentence: ["[CLS]", "The", "doctor", "told", "the", "nurse", "that", "he", "was", "late", ".", "[SEP]"]

- Occupation token: "doctor"
- Extracted embedding: 768-dimensional vector representing "doctor" in this particular sentence context.

These contextual embeddings, capturing subtle linguistic and semantic nuances, form the basis for subsequent feature extraction (Cosine Similarity, PMI, Sentence-Level

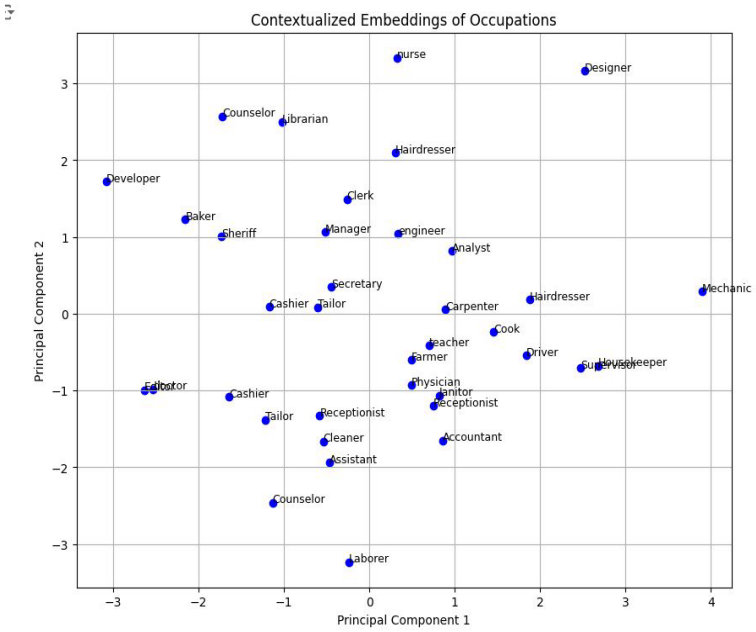

Figure 3:Embeddings of occupations

Contextual Bias).

2.1.4. Feature Extraction and Data Preprocessing

This section thoroughly explains the computational methods and theoretical rationale behind the extraction of the three primary bias features: Cosine Similarity, Pointwise Mutual Information (PMI), and Sentence-Level Contextual Bias. These features serve as foundational elements in constructing the Context-Aware Bias Metric (CABM).

Cosine Similarity (Semantic Bias)

Definition and Justification:

Cosine Similarity measures the semantic similarity between two word embeddings by calculating the cosine of the angle between their corresponding vectors. A high cosine similarity (close to 1) implies that two words share similar semantic contexts, indicating close semantic relatedness. In the context of gender bias, semantic similarity between occupation words and gendered pronouns ("he", "she") helps detect implicit biases that may not be explicitly observable in statistical co-occurrences alone. Thus, Cosine Similarity is essential for quantifying implicit, semantically-driven gender associations within language models.

Method of Computation:

Given two embedding vectors, $\vec{A}$ and $\vec{B}$, the Cosine Similarity is calculated as:

$$Cosine\ Similarity(A,B) = \frac{\overrightarrow{A}.\vec{B}}{\|\vec{A}\|\|\vec{B}\|}$$

In practice, embeddings for gendered pronouns ("he," "she") and occupation terms ("doctor," "nurse") were separately obtained from your fine-tuned BERT model. The Cosine Similarity between each occupation embedding and each gendered pronoun embedding was computed individually. The resulting male-associated and female-associated cosine similarity scores were averaged across their respective pronoun groups, clearly quantifying semantic gender bias for each occupation:

- Male pronouns: "he," "him," "his"
- Female pronouns: "she," "her"

The final semantic bias score for each occupation word is the difference between average male and female cosine similarity scores:

$$Cosine_{bias} = Cos_{male} - Cos_{female}$$

A positive bias score indicates stronger male semantic association, while a negative score indicates stronger female association.

Pointwise Mutual Information (PMI) (Statistical Bias)

Definition and Justification:

PMI quantifies the strength of statistical association between two words based on their co-occurrence frequencies in textual data. PMI is crucial because it explicitly measures how much more often gendered pronouns and occupation terms co-occur than expected by chance. It effectively captures explicit statistical biases present in the language corpus, complementing semantic similarity analysis.

Method of Computation:

The PMI between a pronoun ("he", "she") and occupation ("doctor", "nurse") is defined as:

$$PMI(pronoun, occupation) = \log_2 \frac{P(pronoun,occupation)}{P(pronoun)P(occupation)}$$

Where:

- $P(pronoun, occupation)$ is the joint probability of occupation co-occurring with pronoun.
- $P(pronoun)$ and $P(occupation)$ are individual probabilities of the occupation and pronoun, respectively.

Step-by-step calculation :

- Count occurrences and compute individual probabilities $P(pronoun)$ and $P(occupation)$ for occupations and pronouns, respectively.
- Compute joint probabilities $P(pronoun, occupation)$.

- Calculate PMI scores for each occupation-pronoun pair.
- Aggregate PMI scores for male-associated ("he", "him", "his") and female-associated ("she", "her") pronouns separately.
- The final PMI-based gender bias score is computed as the difference:

$$PMI_{bias(occupation)} = PMI_{male} - PMI\_female$$

Positive PMI bias scores indicate that occupations statistically co-occur more strongly with male pronouns, while negative scores indicate stronger co-occurrence with female pronouns.

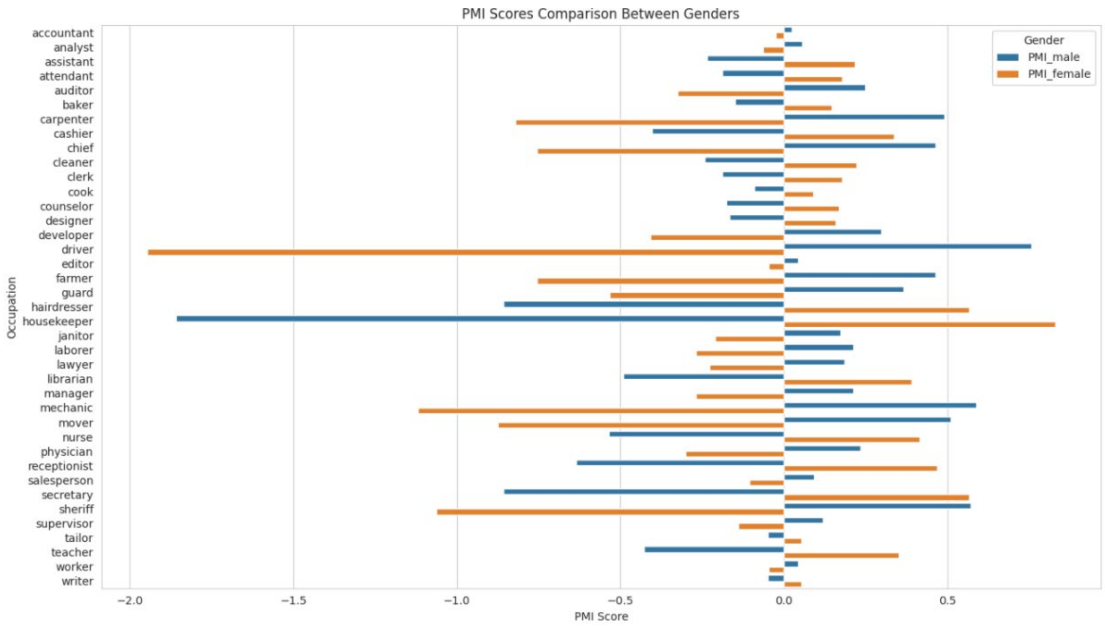

Figure 4: PMI scores between male and female

2.1.5. Sentence-Level Contextual Bias

Definition and Justification:

Transformer-based models like BERT generate context-sensitive embeddings that change depending on how words are used in a sentence. In the case of gendered sentence constructions, this dynamic behavior can lead to subtle but measurable shifts in meaning depending on whether a term appears in a male or female context. The **Sentence-Level Contextual Bias** component of CABM is designed to quantify this variability, reflecting how the overall meaning of a sentence embedding shifts with gendered context.

Unlike traditional word-level metrics, this component works directly with **sentence-level embeddings**, capturing contextual variations that are often invisible to metrics such as PMI or Cosine Similarity alone. This makes it highly suitable for analyzing bias in models like BERT, which rely on full sentence understanding.

Computational Approach:

For each occupation (e.g., "nurse", "engineer"), gendered sentences were extracted from the dataset. These included both male-context sentences (e.g., "He is a nurse.") and female-context sentences (e.g., "She is a nurse.").

- Generate sentence embeddings using the [CLS] token output from a fine-tuned BERT model.
- Compute the **mean embedding** across all male-context sentences and separately for female-context sentences.
- Calculate the **Euclidean Distance** between these two mean embeddings.

$$SentBias(occupation) = \|\vec{e}\_male - \vec{e}\_female\|$$

Where:

- $\vec{e}\_male$: Mean embedding vector for male-context sentences
- $\vec{e}\_female$: Mean embedding vector for female-context sentences

A higher SentBias score indicates greater divergence in how the occupation is represented across gendered sentence contexts, reflecting a stronger contextual bias.

Example:

- Male-context sentence: *"He is a doctor."*
- Female-context sentence: *"She is a doctor."*
- If the distance between their sentence embeddings is high, it suggests that the model understands the same role differently depending on the gendered context.

2.1.6. Composite Metric Design via Feature Weighting Techniques

A key objective of this research is to compute a single, interpretable score that reflects the extent of gender bias present in contextual word embeddings. To achieve this, three independent bias features were extracted: PMI Bias, Cosine Similarity Bias, and Sentence-Level Contextual Bias. However, these features vary in scale, sensitivity, and semantic meaning. Thus, a weighted integration strategy was required to ensure that each component contributes meaningfully to the final score without overpowering the others.

Defining the Context-Aware Bias Metric (CABM)

The Context-Aware Bias Metric (CABM) is a unified metric designed to quantify gender bias using a linear combination of the three extracted features. The general form of the CABM is defined as:

$$CABM(\omega) = \alpha.PMI_{bias(\omega)} + \beta.Cosine_{bias(\omega)} + \gamma.Context_{bias(\omega)}$$

Where:

- $PMI_{bias(\omega)}$: Pointwise Mutual Information bias score for the occupation term w
- $Cosine_{bias(\omega)}$: Cosine similarity bias score for the occupation term w
- $Context_{bias(\omega)}$: Sentence-level contextual bias score for occupation
- α, β, γ: Weighting coefficients assigned to each feature

The accuracy and fairness of the CABM metric depend heavily on how these weights are chosen. To ensure statistical rigor and interpretability, three distinct feature weighting strategies were explored: (1) Principal Component Analysis (PCA), (2) PCA combined with Random Forest Feature Importance, and (3) SHAP-based weights from Random Forest.

2.1.7. 1PCA-Based Feature Weighting

Principal Component Analysis (PCA) is a widely-used unsupervised learning technique that transforms original features into a set of orthogonal components that maximize variance. In this research, PCA was applied to the normalized values of the three extracted features: PMI_Bias, Cosine_Bias, and ContextBias.

Steps:

1. All features were standardized using Z-score normalization.
2. PCA was applied to the scaled feature set.
3. Loadings (weights) from the first principal component were extracted.
4. The absolute values of the loadings were normalized to sum to 1, forming the final weight vector (α, β, γ).

Resulting Formula (example):

$$CABM(\omega) = 0.58 \cdot PMI_{bias(\omega)} + 0.28 \cdot Cosine_{bias(\omega)} + 0.19 \cdot Context_{bias(\omega)}$$

This approach is mathematically grounded and does not rely on labeled data, making it well-suited for unsupervised analysis. However, it does not account for prediction performance or interpretability in real-world tasks.

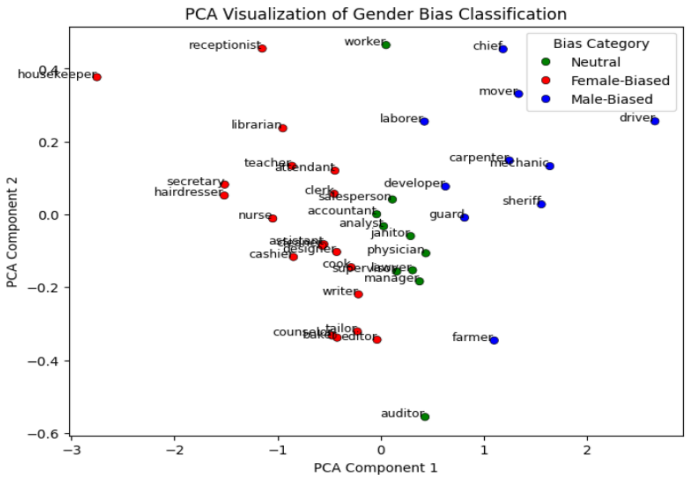

Figure 5:PCA visualization of gender bias classification for each occupations.

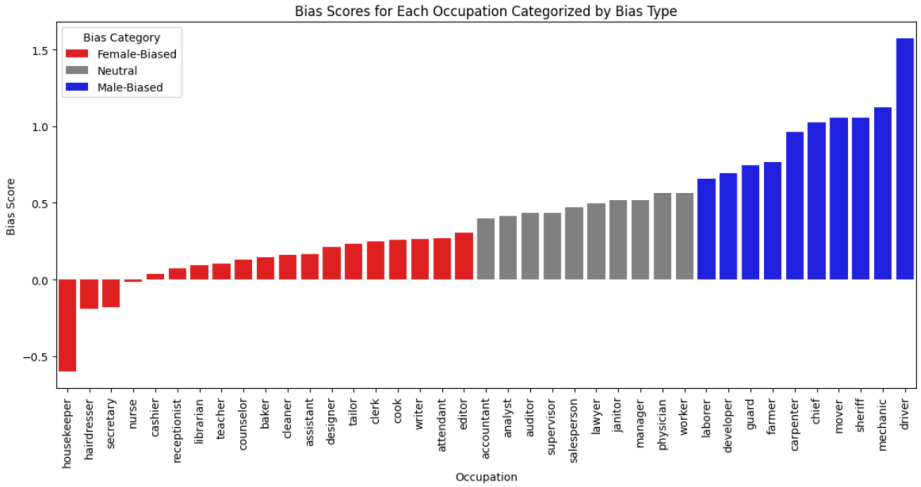

Figure 6: representation of bias score using PCA weights

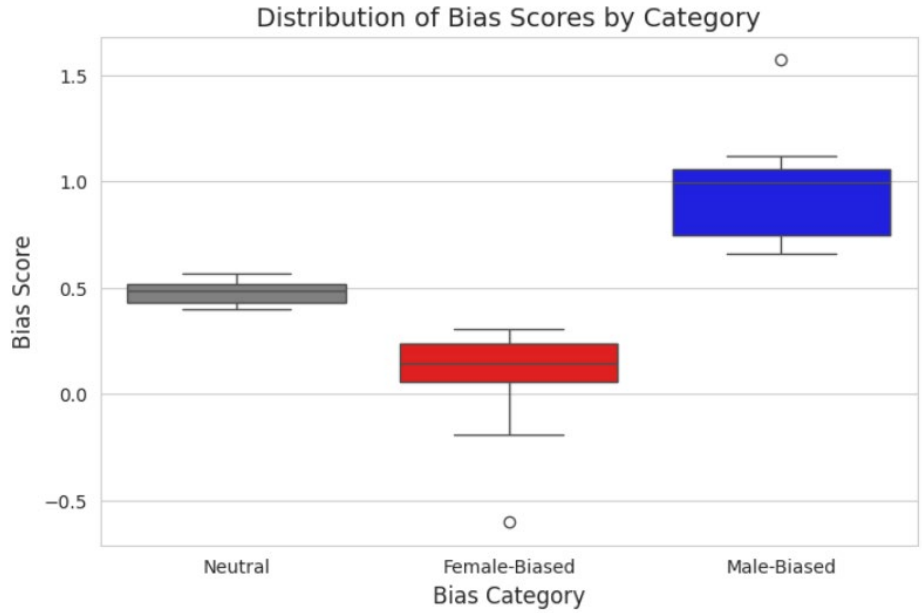

Figure 7: distribution of bias scores by category

2.1.8. PCA + Random Forest-Based Feature Weighting

To combine the benefits of dimensionality reduction and predictive modeling, a hybrid approach was implemented that involved applying PCA followed by a Random Forest model.

Steps:

1. PCA was applied to reduce redundancy and capture the most informative components from the bias features.
2. A **Random Forest Classifier** was trained using the transformed features to predict gender bias categories.
3. **Feature importance scores** were extracted from the trained model.

4. These importance scores were mapped back to the original features and normalized to obtain the final weights.

Resulting Formula (example):

$$CABM\_PCA\_RF(\omega) = 0.42 \cdot PMI_{bias(\omega)} + 0.28 \cdot Cosine_{bias(\omega)} + 0.30 \cdot Context_{bias(\omega)}$$

This method captures non-linear relationships between features while reducing noise through PCA. It improves the reliability of the weights by using prediction accuracy as a basis for feature relevance.

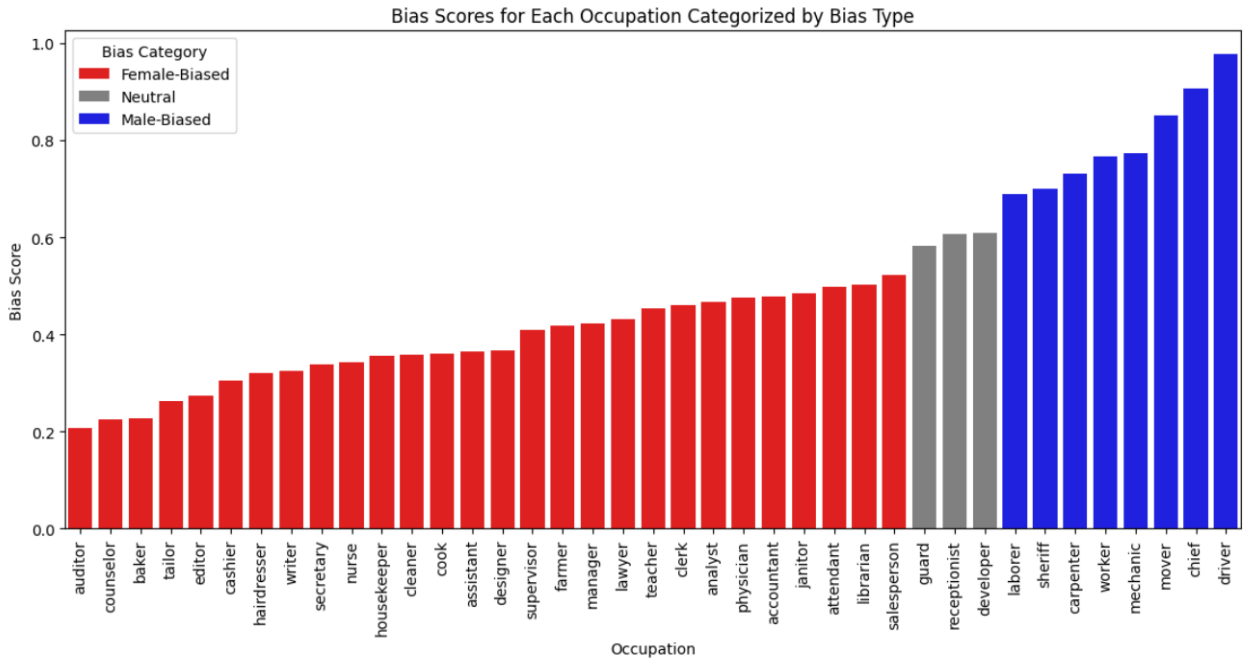

Figure 8:Bias score for each occupation using PCA+RF weighning

2.1.9. SHAP + Random Forest-Based Feature Weighting

The most advanced and interpretable weighting approach used in this research involves combining Random Forest classification with SHAP (SHapley Additive exPlanations) values. SHAP is a game-theoretic technique that provides transparent, localized explanations of each feature's contribution to model predictions.

Steps:

1. A Random Forest model was trained to classify gender-associated bias labels based on the three extracted features.
2. SHAP values were computed for all features to determine their average contribution across the entire dataset.
3. These values were normalized to obtain final weights.

Resulting Formula:

$$CABM\_SHAP(\omega) = 0.583 \cdot PMI_{bias(\omega)} + 0.154 \cdot Cosine_{bias(\omega)} + 0.262 \cdot Context_{bias(\omega)}$$

This approach offers **maximum transparency** and is especially well-suited for explaining why a particular occupation received a high or low bias score. SHAP-based weighting also aligns closely with modern AI fairness standards due to its interpretability and accountability.

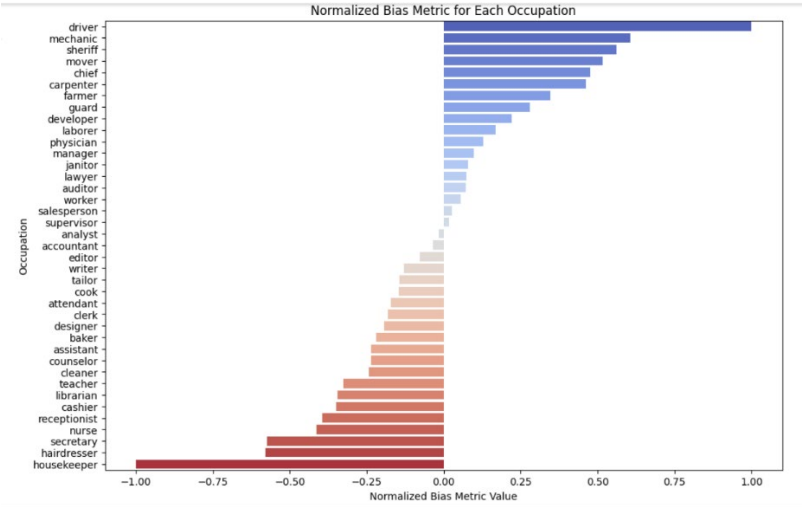

Figure 9: visualization of bias value for each occupation using SHAP+RF weighting.

2.1.10. Metric Validation and Evaluation

This section presents a comprehensive evaluation of the Context-Aware Bias Metric (CABM) across three weighting strategies PCA, PCA + Random Forest, and SHAP + Random Forest. The goal of this evaluation is to ensure that CABM is not only mathematically sound, but also interpretable, consistent, and effective at identifying gender bias in contextual word embeddings.

Distribution Analysis

To examine the distribution of CABM scores across different weighting strategies, we plotted kernel density estimates (KDE) for each method. The results show that SHAP + Random Forest and PCA + Random Forest produce relatively smooth, bell-shaped distributions, suggesting stable and interpretable scoring across occupations. The PCA-only method exhibits slightly more variation, with flatter peaks and broader spread, indicating possible inconsistencies in assigning strong bias scores.

Descriptive statistics confirmed that all three distributions have similar means and standard deviations, with no significant skew. These findings suggest that while all approaches provide usable scores, RF-based methods yield more centralized and distinguishable results, ideal for category-based analysis.

Statistical Significance Testing

To assess the validity of CABM's ability to separate male- and female-biased occupations, we applied two statistical tests:

- **Shapiro-Wilk Test** for normality: All three CABM variants were found to be approximately normally distributed ($p > 0.05$), enabling the use of parametric and non-parametric comparisons.
- **Mann-Whitney U Test**: Comparing SHAP + RF scores between occupations

labeled as “Male-Biased” and “Female-Biased” revealed a statistically significant difference ($U = 198.0$, $p < 0.0001$). This confirms that CABM successfully distinguishes between genders based on embedding behavior and contextual signals, providing strong evidence of its discriminative power.

2.1.11. Correlation Between Weighting Approaches

To measure alignment across methods, both Pearson and Spearman correlation coefficients were computed:

Method Pair	**Pearson (r)**	**Spearman (ρ)**
SHAP + RF ↔ PCA + RF	0.9314	0.9298
SHAP + RF ↔ PCA Only	0.7000	0.6403
PCA + RF ↔ PCA Only	0.7909	0.7047

The high correlation between SHAP + RF and PCA + RF indicates that Random Forest–based methods are consistent, whereas the lower correlation with PCA-only suggests it misses complex patterns in the data. These findings highlight the superior contextual sensitivity of the SHAP-based approach.

2.1.12. Bias Category Agreement and Distribution

To further validate interpretability, occupations were classified into Male-Biased, Female-Biased, or Neutral based on their bias scores. A threshold of +0.1, -0.1 was applied consistently across all three CABM versions. The agreement rates were computed to assess categorical alignment:

Comparison	Agreement (%)
SHAP + RF vs PCA + RF	64.10%
SHAP + RF vs PCA Only	94.87%
PCA + RF vs PCA Only	69.23%

The high agreement between SHAP + RF and PCA Only highlights the consistency of SHAP-driven explainability with traditional dimensionality techniques. PCA + RF, while moderately aligned with PCA Only, diverged more notably from SHAP-based outcomes. Bar chart comparisons of bias category distribution also showed that SHAP + RF provides a directional and explainable classification across Male, Female, and Neutral labels, while PCA + RF leaned conservative, and PCA Only reflected similar patterns with SHAP + RF.

2.1.13. Validation Using Unmasking Predictions

To further validate CABM, we performed unmasking-based pronoun prediction tests using a BERT model fine-tuned on a synthetically biased dataset. This dataset contained controlled gender-occupation distributions (e.g., 90% “he” for engineer, 90% “she” for nurse).

The model was queried with masked sentences like: "[MASK] is a doctor."

Top-1 and Top-2 pronoun predictions and probabilities were extracted. An Unmasking Bias Score was computed as: P(he) - P(she)

- If “he” was predicted with higher confidence → **Male-Biased**
- If “she” was predicted with higher confidence → **Female-Biased**

These unmasking-based bias labels were then compared to CABM bias categories from all three weighting strategies.

SHAP + RF exhibited the highest agreement with the unmasking predictions, confirming its alignment with actual model behavior and supporting its selection as the final weighting strategy.

Weighting Approach	**Total Occupations Compared**	**Matches (Same Bias Category)**	**Accuracy (%)**
SHAP + RF	39	27	69.23%
PCA + RF	39	20	51.28%
PCA Only	39	22	56.41%

2.1.14. Visualization and Explainability

As part of the interpretability framework for the Context-Aware Bias Metric (CABM), several visualization techniques were employed to enhance the transparency of bias measurement outcomes. These visual tools are critical for understanding not just how much bias is present, but also why specific occupations are classified as male-biased, female-biased, or neutral.

2.1.15. SHAP-Based Interpretability

To support explainable decision-making in CABM, the final weighting approach used SHAP (SHapley Additive exPlanations) values derived from a trained Random Forest model. SHAP values provide feature-level attribution, indicating how much each

component PMI, Cosine Similarity, or Contextual Bias contributed to the final bias score of an occupation.

For each occupation, SHAP values were computed and averaged across the dataset. Occupations with strong gender associations (e.g., “nurse,” “engineer”) exhibited clearly distinguishable SHAP contributions, with contextual bias often playing a dominant role. These SHAP values were visualized using summary bar plots, where higher bars represented greater impact on bias classification. Such visualizations offer insight into which features are most responsible for observed bias trends, allowing for targeted evaluation and intervention.

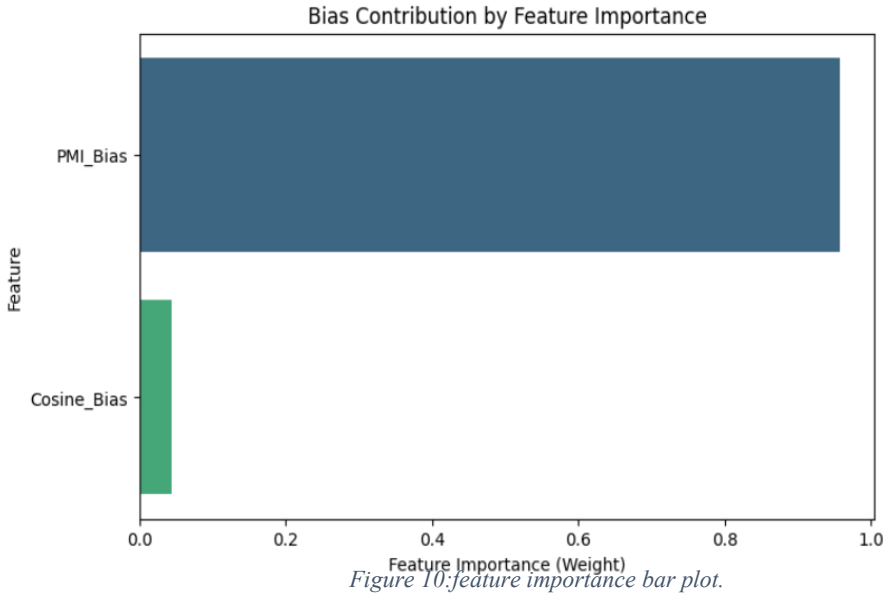

Figure 10:feature importance bar plot.

2.1.16. Heatmap and Bias Score Distribution

To further explore the behavior of CABM across all occupations, a heatmap was generated to visualize the interaction between gender (male vs female) and bias scores. This visualization highlighted:

- Which occupations were consistently biased across all weighting strategies
- The magnitude and direction of these biases
- Patterns in bias alignment between PCA, PCA + RF, and SHAP + RF variants

Additionally, KDE (kernel density estimate) plots were used to assess the distribution of bias scores across the three CABM methods. These plots provided a smoothed view of score density, revealing whether bias scores were skewed, centered around neutrality, or polarized toward one gender. SHAP + RF demonstrated a balanced and stable distribution, further justifying its selection as the primary method.

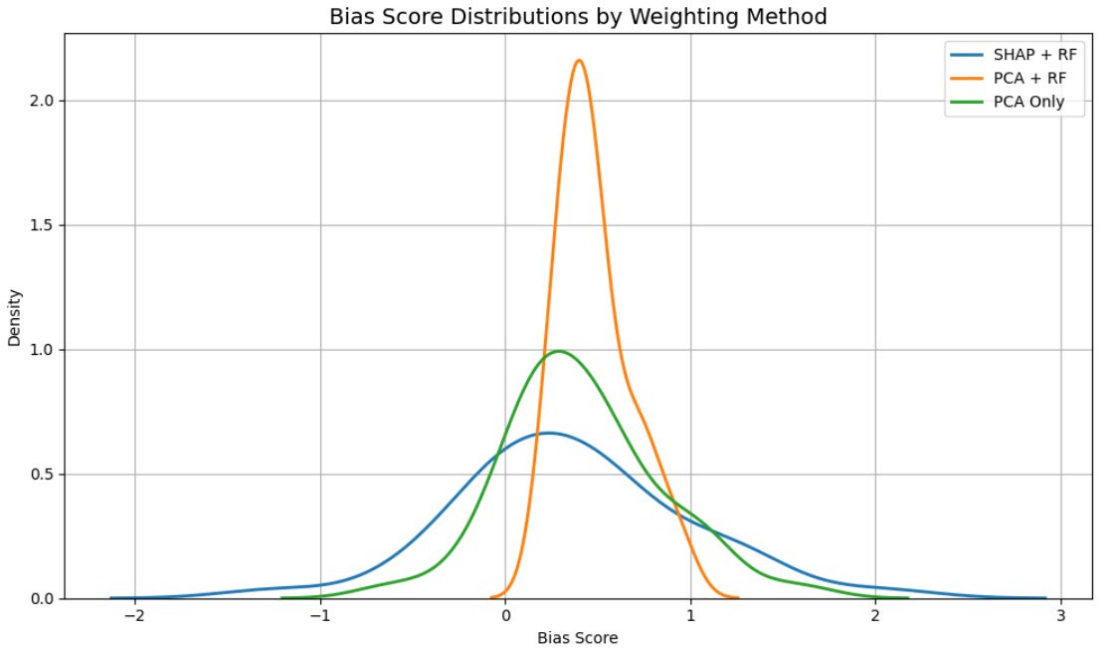

Figure 11:Kernel Density Estimation (KDE) plots showing the distribution of bias scores across three weighting strategies used in CABM: SHAP + RF, PCA + RF, and PCA only. SHAP-based scoring shows a smooth, centralized distribution, supporting its selection as the most stable and interpretable approach.

2.1.17. Bias Category Distribution and Agreement Visualization

To evaluate how different weighting methods classified occupations, bar charts were created to show the number of occupations categorized as Male-Biased, Female-Biased, or Neutral. These visualizations revealed:

- SHAP + RF exhibited a directional and interpretable distribution, capturing both

Female-Biased and Male-Biased occupations with minimal neutral assignments.

- PCA + RF showed a conservative tendency, assigning the highest number of occupations as Neutral, with fewer Female-Biased labels.
- PCA Only identified the most Female-Biased occupations and showed a distribution that aligned closely with SHAP-based classifications.

In addition, category agreement matrices and percentage agreement calculations were used to show how often two methods agreed on the same bias label for an occupation. SHAP + RF and PCA Only exhibited the highest agreement (94.87%), while PCA + RF showed weaker alignment, especially with SHAP-based classifications.

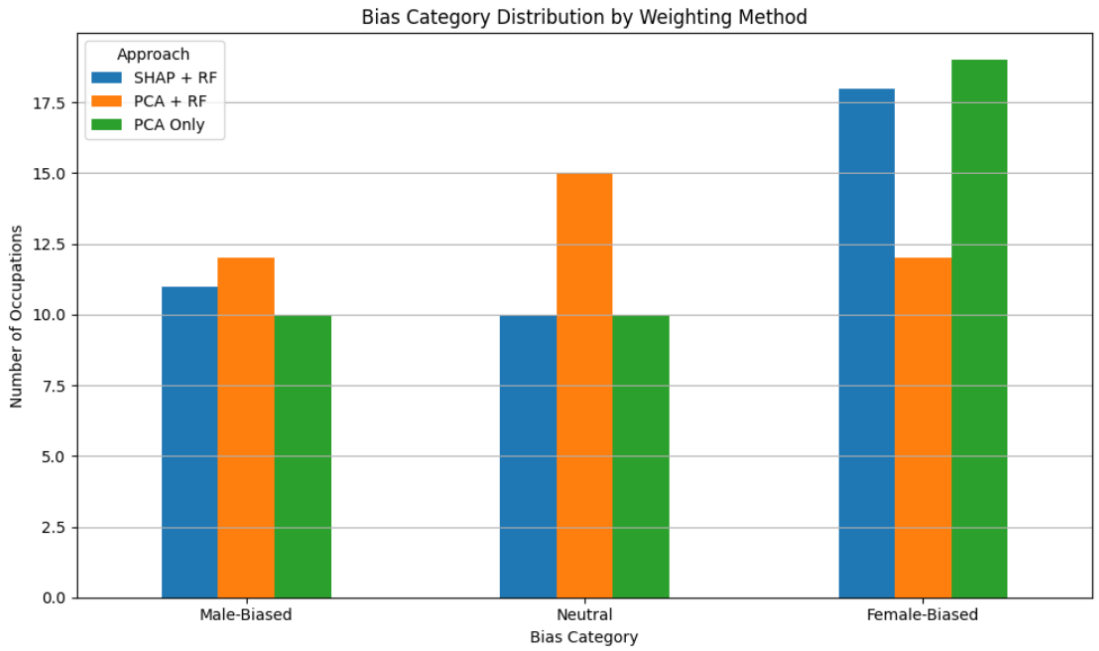

Figure 12:Bias category distribution across occupations for each CABM weighting method, based on dataset-labeled categories. SHAP + RF and PCA + RF show a more balanced distribution across Male-, Female-, and Neutral-biased occupations, while PCA Only identifies more Female-Biased roles. The results demonstrate how different weighting strategies influence bias classification outcomes.

2.1.18. Embedding Similarity Visualization

To support sentence-level contextual analysis, embedding similarity plots were generated to illustrate how the same occupation word shifted in BERT's embedding space across gendered contexts. For example, the word *"doctor"* had slightly diverging embeddings in

"He is a doctor" versus *"She is a doctor"*. These shifts were quantified using cosine similarity or Euclidean distance, and plotted to show occupations with the highest contextual variation one of CABM's key components.

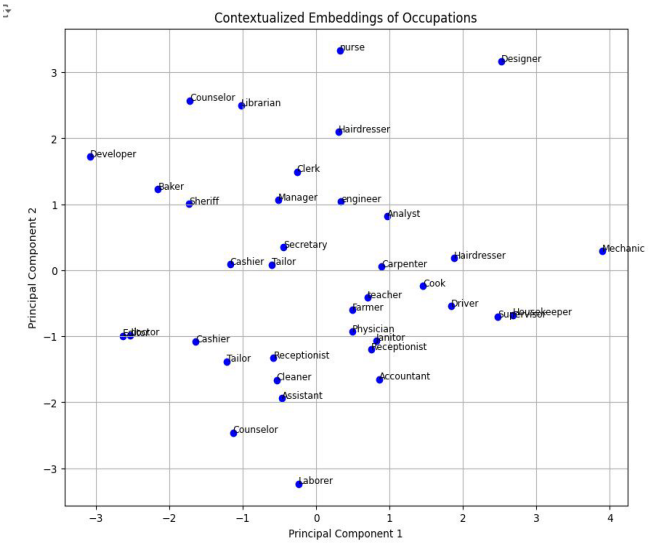

Figure 13: embeddings of each occupation

2.1.19. Reproducibility and Extensibility

Tools and Frameworks

The following tools and frameworks were used throughout the implementation:

- Python 3.10: Primary programming environment.
- Hugging Face Transformers: Used for embedding extraction from BERT.
- Scikit-learn: Employed for PCA, Random Forest training, and statistical evaluations.
- SHAP: Used for explainable AI feature attribution and visualization.
- Pandas & Seaborn/Matplotlib: For data handling and visual analysis.
- Google Colab: Provided a scalable and cloud-accessible platform for running all computations and visualizations.

All code was executed in a Google Colab environment with GPU acceleration (when available), and relevant outputs (plots, CSVs, and model explanations) were saved to Google Drive for documentation and reproducibility.

2.1.20. Modular Design for Reproducibility

Each phase of the metric computation was encapsulated in a dedicated module, including:

1. Data Preprocessing: Tokenization, gender labeling, and masked sentence creation.
2. Embedding Extraction: BERT-based contextual embedding computation.
3. Bias Feature Extraction: PMI calculation, Cosine similarity, and Contextual Bias score.
4. Metric Calculation: CABM formula application with user-defined or learned weights.
5. Validation & Analysis: Includes statistical tests, category assignment, and SHAP-based explanation.

This modular design allows researchers to independently test or replace components (e.g., switch BERT with RoBERTa, or compare SHAP with LIME for interpretability).

2.1.21. Extending to Other Models and Bias Domains

CABM was designed to be model-agnostic. While BERT was used for this study, any contextual embedding model such as RoBERTa, ALBERT, or GPT can be substituted with minimal adjustments.

Similarly, while this study focuses on **gender bias**, the CABM framework can be extended to other dimensions of bias such as:

- **Racial bias** (e.g., comparing associations with Black vs. White identity terms)
- **Age bias** (e.g., analyzing terms like “elderly” vs “young”)

- **Intersectional bias**, as demonstrated in recent work on multi-class identity associations

2.1.22. Testing and Implementation

The implementation and testing of the Context-Aware Bias Metric (CABM) were carried out through a modular and reproducible pipeline. This section outlines the development environment, implementation stages, evaluation processes, and additional experiments conducted to explore the metric's sensitivity to synthetically injected bias.

Development Environment

All components of CABM were implemented in Python 3.10 using the Google Colab platform with GPU support. The following libraries were used:

- Transformers (Hugging Face): BERT embedding extraction and MLM fine-tuning
- Scikit-learn: PCA, Random Forest modeling, statistical analysis
- SHAP: Feature-level explanation for Random Forest models
- Matplotlib & Seaborn: Data visualization (KDE, bar plots, SHAP graphs)
- Pandas, NumPy: Data handling and feature computation
- SciPy: Statistical testing (e.g., Mann-Whitney U, Shapiro-Wilk, correlation tests)

CABM Pipeline Implementation

The pipeline was structured into modular phases:

1. Data Preparation: Filtering and annotating sentences from the WinoBias dataset, resulting in 927 contextual sentences with gender labels and occupation tags.
2. Embedding Extraction: Used bert-base-uncased to compute [CLS] and token embeddings from masked and unmasked sentences.
3. Feature Computation:
    - Cosine Similarity between gendered pronouns and occupations
    - Pointwise Mutual Information (PMI) based on co-occurrence

    - Contextual Bias Score via embedding shift across gendered sentence frames
4. Metric Computation:
    - Combined features using three weighting strategies: PCA, PCA + RF, SHAP + RF
    - Computed final CABM scores and classified occupations
5. Validation:
    - Performed statistical testing (normality, significance, correlation)
    - Computed agreement rates and category distributions
    - Evaluated visual interpretability using SHAP and other plots
    - Validation Using Unmasking Predictions

2.1.23. Additional Testing: Synthetic Bias Injection via MLM Fine-Tuning

To further validate the robustness of CABM, an experiment was conducted to assess whether the metric could detect bias injected into a language model via training. A custom biased dataset was created, where each occupation was synthetically linked with a preferred gender:

- For example, *engineers* were paired with "he" in 90% of cases, while *nurses* were paired with "she" in 90% of cases.

Figure 14:example of dataset creation prob

This dataset was used to fine-tune a BERT model using the Masked Language Modeling (MLM) task. The intention was to train the model to learn gender-stereotypical associations in a controlled setting.

2.1.24. Unmasking Tests for Evaluation

After fine-tuning, test sentences were constructed with masked pronouns, such as:

Sentence: “The janitor cleaned the office after [MASK] left the building.”

The model was prompted to predict the pronoun at the masked location. Example outputs:

→ the janitor cleaned the office after he left the building. (score: 0.6349)

→ the janitor cleaned the office after she left the building. (score: 0.3630)

→ the janitor cleaned the office after they left the building. (score: 0.0005)

2.1.25. Metric Evaluation on Fine-Tuned Model

To empirically validate the reliability of the Context-Aware Bias Metric (CABM), we designed a controlled fine-tuning experiment using three distinct BERT models:

1. **Biased BERT** – fine-tuned on a custom synthetic dataset (7928) where occupations were associated with predefined gender distributions (e.g., "engineer" paired with "he" 90% of the time).
2. **Balanced BERT** – fine-tuned on a balanced dataset (8000) with equal male/female sentence representation for all occupations (50:50 distribution).
3. **Normal BERT** – the original pretrained bert-base-uncased model without fine-tuning.

The objective was to determine whether CABM can detect and reflect the presence or absence of gender bias learned during training. After fine-tuning, each model was evaluated using two parallel techniques:

- **Unmasking predictions**, to reveal the model's behavioral bias.
- **CABM scoring**, to measure embedding-level gender bias using our proposed metric.

Each model was prompted with masked gender-neutral sentences (e.g., "[MASK] is an engineer"). The predicted pronouns (e.g., "he" or "she") and their associated probabilities were extracted. In the biased model, stereotypical pronouns dominated top-1 predictions (e.g., "he" = 0.876 for "engineer"). The balanced model produced more even predictions, and the normal model reflected moderate societal bias.

This controlled three-model comparison demonstrates that CABM is not only sensitive to training-induced bias but also **mirrors real model behavior** observed through unmasking predictions. The experiment confirms CABM's effectiveness as a reliable and interpretable tool for quantifying gender bias in contextual word embeddings.

2.1.26. Comparability of Word-Wise Biases

To evaluate the robustness of CABM, a comparison was made with the established Word Embedding Association Test (WEAT). Following the method of Schröder et al. [24], both

metrics were tested across three comparable BERT models: one pretrained, one fine-tuned on a synthetically biased dataset, and one fine-tuned on a balanced dataset. All three models were evaluated on the same set of occupation-related sentences. For each occupation, the standard deviation of scores across models was calculated to assess how consistently each metric quantifies bias.

WEAT achieved a lower average standard deviation (0.0111), showing high numerical stability. CABM, while slightly more variable (0.0385), remained consistent and added richer context awareness due to its use of sentence-level embeddings.

Overall, although WEAT is more stable, it lacks sensitivity to contextual nuances. CABM balances reliability with interpretability, making it a more expressive and modern bias detection metric.

WEAT	0.0111	High stability but limited context
CABM	0.0385	Context-aware and statistically robust

2.1.27. Insights and Limitations

- This testing approach demonstrated how CABM could be applied to evaluate contextual bias introduced via training.
- The unmasking predictions gave interpretable evidence of potential pronoun preferences.
- However, due to the limited dataset size, the BERT model did not generalize the bias patterns effectively.
- Future iterations will require a larger, balanced, and diverse biased dataset to produce more conclusive results.

2.1.28. Commercialization Aspects of the Metric

The Context-Aware Bias Metric (CABM) developed in this research demonstrates high potential for commercialization as a modular, interpretable, and extensible framework for auditing gender bias in language models. As the use of contextual word embeddings continues to expand in critical domains such as hiring, education, law, healthcare, and public policy, the demand for transparent and robust bias detection tools is growing. CABM addresses this need by offering a practical solution that captures semantic, statistical, and contextual dimensions of bias in transformer-based NLP systems.

Designed with industry integration in mind, CABM is compatible with widely-used machine learning frameworks (e.g., Scikit-learn, Hugging Face Transformers, SHAP) and can be seamlessly embedded into both academic research workflows and enterprise NLP pipelines. Its modular architecture enables organizations to adapt the metric to different embedding models (e.g., RoBERTa, ALBERT, GPT) or to extend it to other bias dimensions beyond gender, such as race, age, or intersectionality.

CABM can be productized and deployed in various formats, including:

- A Python-based toolkit for model developers and AI fairness researchers
- A command-line utility or CI/CD plug-in for continuous bias auditing
- A web-based dashboard offering bias visualizations, comparisons across models, and downloadable reports

In commercial environments, CABM could serve as a core component of responsible AI platforms and algorithmic auditing systems. Companies deploying large language models could integrate CABM during pre-deployment validation phases to ensure alignment with fairness regulations and ethical standards. Its explainability layer, powered by SHAP, allows both technical and non-technical stakeholders to understand how and why a model's outputs may be biased an essential requirement under emerging transparency laws (e.g., GDPR, AI Act).

Furthermore, CABM's visual output modules including KDE plots, category bar charts, SHAP summaries, and contextual embedding comparisons support effective communication with multidisciplinary teams, including data scientists, product managers, and policy analysts.

In summary, CABM stands out as a commercially viable and academically rigorous framework for bias diagnosis in contextual language models, offering transparency, flexibility, and interpretability. Its implementation fills a critical gap in the AI development lifecycle, where fairness evaluation must go beyond surface-level metrics and engage directly with contextual meaning in language.

2.2. Developing Metrics for Detecting Gender Bias in Image Datasets Using Contextual Factors: Objects, Scenes, And Spatial Relationships

2.2.1. Overview of the Research Framework

This study proposes a structured, multi-stage methodology to detect and quantify contextual gender bias in visual datasets through a novel metric called the Unified Bias Metric (UBM). The framework is designed to measure how both object-level features (e.g., size, spatial positioning, depth) and scene-level semantics (e.g., background context) influence AI-driven gender classification outcomes. In contrast to traditional fairness evaluations that rely on co-occurrence or demographic group statistics, this approach

captures contextual dependencies by analyzing how individuals and objects are visually arranged within each scene.

The UBM pipeline consists of four main stages: data preparation, feature extraction, bias scoring, and interpretability analysis. Its design prioritizes scalability, reproducibility, and extensibility. At its core, the metric integrates two principal components:

- **Object Influence Score (OIS):** Quantifies the influence of object-level spatial features including relative size, 3D proximity, and depth difference on gender classification.
- **Scene Similarity Bias (SSB):** Measures the semantic alignment of image scenes with male- or female-dominant contexts using pre-trained scene embeddings.

The final contextual bias score is calculated as a weighted combination of OIS and SSB:

$$UBM = \alpha * OIS + \beta * SSB$$

where the weights (α, β) are obtained through dimensionality reduction techniques (e.g., PCA), explainability models (e.g., SHAP), or predictive regression models (e.g., Ridge).

Framework Pipeline Overview

The complete pipeline consists of the following steps:

1. Dataset Curation and Gender Labeling

A filtered subset of the Microsoft COCO dataset [25] is selected, focusing on images where a person co-occurs with objects commonly associated with gender stereotypes. Binary gender labels are manually annotated for the most prominent person in each image.

2. Object Detection and Segmentation

Persons and objects are detected using YOLOv8 (Ultralytics, 2023) [26]. For improved segmentation precision, masks are refined using the Segment Anything Model (SAM) [27]. Bounding boxes, masks, and object labels are stored in JSON metadata for downstream analysis.

3. Depth Estimation and 3D Contextual Analysis

Depth maps are generated using the DepthAnything v2 model [28], enabling the computation of normalized object-person distances in both 2D and 3D space. These features feed into the OIS component.

4. Scene Embedding and Contextual Bias Evaluation

Semantic embeddings are extracted using CLIP (OpenAI, 2021) [29] and optionally supported with Places365 [30] scene classifications. Cosine similarity is computed between image scene embeddings and gender-reference scene vectors to produce the SSB.

5. Feature Normalization and Engineering

All extracted features relative size, 3D distance, depth, and scene similarity are normalized to ensure comparability across images of varying scale and resolution.

6. UBM Score Computation and Weight Optimization

The UBM score is computed using multiple weighting strategies across nine experimental approaches:

- PCA-based variance weighting
- SHAP-based feature attribution
- XGBoost and Ridge regression modeling
- Optuna-optimized model tuning

These variations allow for evaluating robustness and sensitivity under different learning assumptions.

7. Interpretability and Visualization

Visual outputs include:

- SHAP plots for feature influence
- Bar and violin plots for object-level bias scores

- Scene embedding similarity maps

All results and visualizations are generated using Python (matplotlib, seaborn) and stored as CSV files for reproducibility.

Figure 15: UBM pipeline

This methodology offers a modular, extensible, and reproducible pipeline for measuring contextual bias in AI vision systems. By combining explainable machine learning, semantic embeddings, and spatial modeling, UBM provides a tool for auditing gender bias in image datasets across diverse domains such as surveillance, hiring, and content generation.

2.2.2. Dataset Preparation

The effectiveness of the Unified Bias Metric (UBM) relies heavily on the quality and structure of the dataset used to capture contextual gender bias. To ensure that the metric evaluates images with meaningful object-person interactions and contextual diversity, a curated subset of the Microsoft COCO dataset [25] was selected. The preparation process

involved systematic filtering, manual labeling, and metadata structuring, with an emphasis on gender-associated object co-occurrences.

Dataset Selection

The Microsoft COCO dataset was selected due to its comprehensive annotations, diversity of visual contexts, and the availability of object and person instances in a wide range of everyday scenes. For this research, the filtering criterion required that each image must contain at least one person to study how visual context (e.g., objects, background) might influence gender classification. This subset is especially valuable for analyzing contextual biases, as it reflects real-world co-occurrences between people and objects in various environments.

Object Category Filtering

For target potential contextual gender biases, eight object categories were selected based on prior research [4][5][6] and sociocultural stereotypes. These categories were divided as follows:

- Female-associated objects: handbag, hair drier, umbrella, cup
- Male-associated objects: sports ball, baseball bat, bicycle, skateboard

Using the COCO annotations file (instances_val2017.json), a custom Python script filtered out images that featured at least one person and one of the above objects. This step ensured that the final dataset reflected a balance of male- and female-associated contexts likely to trigger bias in AI vision models.

Manual Gender Labeling of Persons

Because the COCO dataset does not include gender annotations, manual labeling was performed. The labeling focused on the most prominent person in each image defined by the largest bounding box area and central positioning.

To improve reliability:

- Two annotators independently assigned binary gender labels (male/female).
- Inconsistencies were resolved through discussion or exclusion.
- Only one gender label was assigned per image to ensure consistent analysis.

These gender labels served as ground truth for evaluating object-level bias in the UBM pipeline.

Metadata Structuring

For each image, a structured JSON file was created to store metadata, which included:

- Image ID and file name
- Bounding boxes and class names of detected objects
- Bounding box of the labeled person
- Manually annotated gender label

Placeholders for computed features such as object size, spatial distance, depth, and scene similarity.


```
    {
        "class_name": "person",
        "bbox": [
            172.33811950683594,
            156.70892333984375,
            481.0,
            637.9450073242188
        ],
        "confidence": 0.8721827268600464,
        "mask_path":
"/content/drive/MyDrive/ContextualBiasProject/outputs/all_images/masks/000000000036/mask_person_1.png",
        "mean_depth": 130.99079587429102,
        "normalized_depth": 51.3689395585455,
        "gender": "Female"
    },
    {
        "class_name": "umbrella",
        "bbox": [
            5.67034912109375,
            56.59185791015625,
            453.52667236328125,
            526.848876953125
        ],
        "confidence": 0.8918861746788025,
        "mask_path":
"/content/drive/MyDrive/ContextualBiasProject/outputs/all_images/masks/000000000036/mask_umbrella_0.png",
        "mean_depth": 118.64993351063829,
        "normalized_depth": 46.52938569044639,
        "relative_size": 1.41785825160928,
        "normalized_distance_xy": 0.000810992856981018,
        "normalized_distance_z": 25.44449020520048,
        "3d_distance": 25.444490218124876
```


Figure 16:Example JSON Metadata Format.

Final Dataset Composition

After filtering and annotation, the final dataset included **852 images**, evenly distributed across the selected object categories. Gender labels were approximately balanced, ensuring meaningful bias detection across diverse visual contexts. This finalized dataset was passed through the UBM pipeline for feature extraction, contextual analysis, and metric computation.

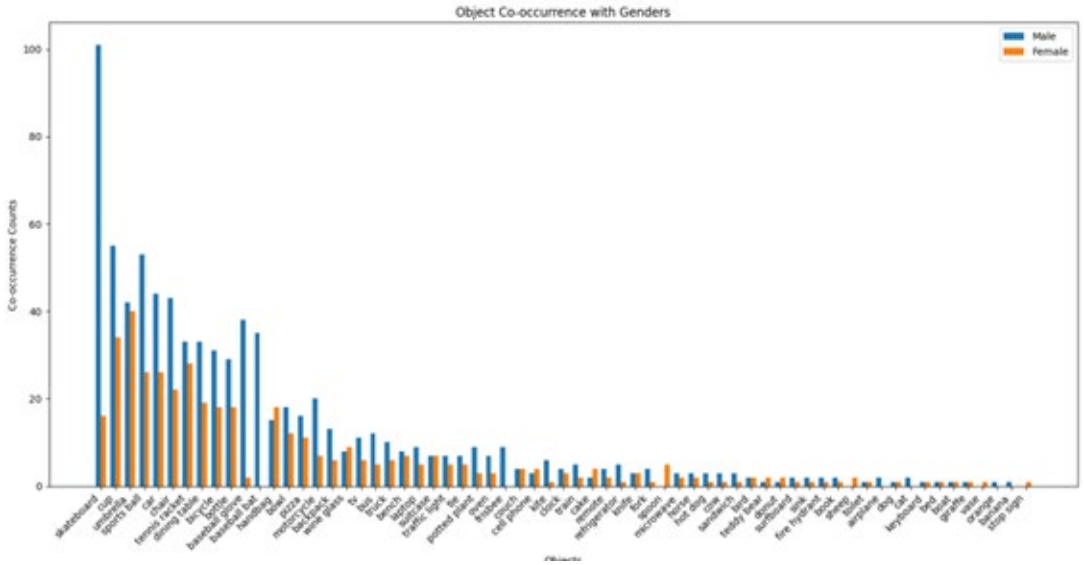

Figure 17:Bar Chart of Object Occurrences by Gender

2.2.3. Object Detection and Segmentation

The accurate detection and segmentation of persons and objects within each image is a foundational step in the Unified Bias Metric (UBM) pipeline. This phase ensures that all visually relevant elements particularly those associated with gender stereotypes are identified, localized, and structured for further analysis such as object-person spatial relationships and feature extraction. High-quality object detection directly influences the reliability of the Object Influence Score (OIS), a critical subcomponent of the UBM.

Model Selection & Tools

To achieve robust and precise detection, two state-of-the-art computer vision models were integrated into the pipeline:

- YOLOv8 (You Only Look Once, Version 8) [26]: A real-time, high-performance object detection model by Ultralytics, used to detect all COCO-labeled objects including persons. YOLOv8 was selected for its lightweight architecture, superior inference speed, and high accuracy across varied object categories.
- Segment Anything Model (SAM) [27]: An advanced segmentation model used optionally to enhance YOLO-generated bounding boxes by producing fine-grained, pixel-wise masks. This refinement step is especially useful for irregular or overlapping objects (e.g., umbrellas, cups), providing more accurate area estimations necessary for size-based bias computation.

Detection and Segmentation Process

Each image is first passed through YOLOv8 to extract:

- Bounding boxes for all persons and detected objects.
- Class labels based on COCO object taxonomy.
- Confidence scores, filtered using a threshold of ≥ 0.25.

For segmentation refinement, detected bounding boxes are fed into SAM, which outputs pixel-wise masks for improved spatial delineation. These masks are then linked to the corresponding object metadata.

Metadata for each object includes:

Table 1:Sample YOLO + SAM Output Metadata

Field	Description
Class Label	COCO category (eg : handbag, person)
Bounding Box	Coordinates(x, y, width, height)
Area	Calculated from mask or bounding box
Confidence Score	Detection confidence (range : 0 – 1)
Mask ID	Index for linking to segmentation output

Relative Object Size Calculation

For each detected object, a **relative size score** is computed to reflect its visual prominence. This is calculated as the ratio of the object's bounding box area to the full image area. If segmentation masks from SAM are available, the pixel-wise mask area is used instead for higher accuracy, especially for irregularly shaped or partially occluded objects.

All relative size values are **normalized** across the dataset to maintain consistency and enable comparative analysis across images of varying dimensions. This feature is crucial because larger or foreground objects tend to dominate the visual field and may disproportionately influence gender predictions particularly when such objects are stereotypically gendered (e.g., handbags, sports balls, hair dryers). Relative size is one of the three core components used in computing the **Object Influence Score (OIS)**.

Identification of Prominent Person

To standardize gender annotation and contextual analysis, only the most prominent person per image is retained. Prominence is determined using:

- Largest bounding box area, and
- Proximity to image center (Euclidean distance)

This approach ensures consistency in measuring object influence relative to a single subject per image.

Output & Storage Format

Each processed image produces a dedicated metadata JSON file, storing:

- Image ID and filename
- Bounding boxes and labels of all detected objects
- Bounding box and center-point of the main person
- Optional segmentation mask references (if SAM is enabled)

Figure 18: YOLOv8 Detection Output

Figure 19: SAM segmentation Image

Limitations of Segmentation Models

While SAM significantly improves spatial granularity, its performance may degrade in cases involving:

- Heavily occluded objects
- Overlapping object masks
- Low-contrast scenes or ambiguous boundaries

Acknowledging these limitations allows for nuanced interpretation of object influence, particularly when comparing between bounding-box-only and refined-mask pipelines.

This module ensures that all gender-relevant visual cues are captured with high fidelity, providing the structural input for downstream spatial, depth, and contextual bias analysis.

Together, YOLOv8 and SAM create a hybrid detection system that balances speed, accuracy, and segmentation detail, setting the foundation for explainable and interpretable bias quantification.

2.2.4. Depth Map Generation

Depth estimation is a crucial component of the Unified Bias Metric (UBM) framework, particularly for computing the Object Influence Score (OIS). While traditional 2D object features like size and position offer insight into visual prominence, incorporating depth enables a more accurate representation of how close or far an object is from the subject (person) in 3D space. This is especially important for evaluating contextual bias, as foreground objects often carry more visual weight in classification systems.

Importance of Depth in Contextual Bias Detection

In both human and machine vision, objects appearing closer are typically more salient and exert a stronger influence on perception. In the context of gender bias, for example, a nearby object such as a handbag or sports ball may significantly skew a model's gender prediction, while a similar object in the background may have negligible effect.

Integrating depth into the UBM pipeline:

- Enhances contextual proximity detection
- Distinguishes influential cues from background noise
- Increases interpretability and reliability of bias scores

Depth Estimation Using DepthAnything v2

To achieve accurate depth prediction, the study employs DepthAnything v2, a state-of-the-art monocular depth estimation model [28]. Trained on large-scale unlabeled datasets, it produces smooth, high-resolution grayscale depth maps from a single RGB input.

- Brighter regions: Indicate closer objects
- Darker regions: Represent distant background

Each image in the dataset is processed through this model, generating depth maps stored alongside original images for downstream analysis.

Figure 20: Original Image vs. Generated Depth Map

Depth Feature Extraction

Once the depth maps are generated, depth-based features are extracted for each detected object using its mask or bounding box. The key features include:

- Mean depth of the object (D_{object})
- Mean depth of the person (D_{Person})
- Z-distance (relative depth):

$$Z_{distance} = D_{object} - D_{Person}$$

All values are normalized to a [0, 1] range to ensure comparability across lighting conditions and contrast variances.

Figure 21:Overlay of Object Masks with Depth Values

3D Distance Calculation

To compute the spatial relationship between each object and the person in 3D space, the full **Euclidean distance** is calculated using:

$$3D\ Distance = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2}$$

Where:

- (x_1, y_1): Center of the person's bounding box
- (x_2, y_2): Center of the object's bounding box
- (z_1, z_2): Mean depth values for the person and object respectively

This 3D distance is a critical component of OIS, helping distinguish between contextually important foreground and less relevant background elements.

Integration with the UBM Pipeline

All depth-related features are added to each image's metadata (in JSON format) and then passed to:

- Object Influence Score (OIS) computation
- UBM score calculation (OIS + SSB)

- Explainability models (e.g., SHAP, PCA)

This integration ensures spatial awareness is incorporated at every level of the pipeline.

By including high-fidelity depth estimation via DepthAnything v2, the UBM framework benefits from improved 3D modeling of object-person interactions. This added dimensionality enhances the robustness, fairness, and interpretability of contextual bias measurement in visual AI systems.

2.2.5. Scene-Level Feature Extraction

While object-level features offer insight into localized bias (e.g., object size or proximity to the subject), scene-level context captures global semantics such as environment, spatial arrangement, and lighting conditions. These latent cues often ignored by traditional bias audits can significantly influence AI gender classification outcomes. In the Unified Bias Metric (UBM) framework, this broader context is captured through the **Scene Similarity Bias (SSB)**: a score that reflects how semantically "masculine" or "feminine" a scene appears based on pre-trained vision-language associations.

Scene Embedding Using CLIP

To extract high-level scene semantics, each image is passed through CLIP (Contrastive Language–Image Pretraining) [3], a multi-modal model trained on 400M image-text pairs. CLIP encodes each image into a 512-dimensional embedding vector $\boldsymbol{E_{image}} \in \mathbb{R}^{\mathbf{512}}$, which captures global visual features—layout, textures, lighting, and compositional structure without requiring explicit object segmentation.

CLIP was selected for:

- Robust generalization across diverse domains
- Pre-training on large-scale paired image-text data
- Compatibility with cosine similarity-based comparison

These embeddings serve as the foundation for scene-gender alignment analysis.

2.2.6. Construction of Gendered Scene Embedding References

To compute SSB, two reference vectors are built by averaging the CLIP embeddings of hand-selected gender-associated scenes:

- $\boldsymbol{E_{male}}$: Mean embedding of scenes stereotypically associated with males. (e.g., **garage**, **stadium**, **gym**, **workshop**)
- $\boldsymbol{E_{female}}$: Mean embedding of scenes associated with females. (e.g., **kitchen**, **dressing room**, **nursery**, **living room**)

Formally:

$$E_{male} = \frac{1}{N}\sum_{i=1}^{N} CLIP(image_{male}, i) \ , E_{female} = \frac{1}{M}\sum_{j=1}^{M} CLIP(image_{male}, j)$$

These reference vectors provide semantic anchors for gender-context comparison.

Scene Similarity Bias (SSB) Calculation

Each image's embedding is compared against both reference embeddings using cosine similarity:

$$SSB = \cos(E_{image}, \quad E_{male}) - (E_{image}, \quad E_{female})$$

Where:

- $SSB > 0$: Scene is more similar to male-associated environments
- $SSB < 0$: Scene aligns more with female-associated environments
- $SSB \approx 0$: Scene is contextually neutral

The computed SSB is saved in the image's metadata and combined with OIS during final UBM computation:

$$UBM = \alpha \times OIS + \beta \times SSB$$

Where α and β are weights derived via PCA, SHAP, or regression tuning strategies.

Scene-level feature extraction is crucial to capturing latent contextual bias that may not be reflected in objects alone. The Scene Similarity Bias (SSB), computed via CLIP embeddings, enables UBM to quantify global environmental influence. By comparing scene semantics against reference vectors, this method delivers an explainable and reproducible bias score that extends the fairness analysis beyond subject-centric methods.

2.2.7. Unified Bias Metric (UBM) Formulation

The Unified Bias Metric (UBM) is the central formulation of this research, designed to quantify contextual gender bias in images by integrating both object-level and scene-level features. Unlike traditional metrics that rely solely on demographic co-occurrence or class imbalance, UBM captures how visual context through object salience, spatial dynamics, and semantic background influences gender classification outcomes. The metric is computed per image and can be aggregated across object categories or datasets for large-scale bias audits.

UBM consists of two core components:

- **Object Influence Score (OIS)** — Captures the spatial prominence and salience of gendered objects relative to the person.
- **Scene Similarity Bias (SSB)** — Captures semantic similarity between the scene and gender-associated environments.

Object Influence Score (OIS)

The Object Influence Score (OIS) represents the degree to which a particular object is visually dominant or likely to influence model perception of the subject. It is computed as a weighted combination of three normalized features:

- **Relative Size (S):** Ratio of object area to image area (or from segmentation mask)
- **3D Distance (D):** Euclidean distance from the object to the person using spatial + depth coordinates
- **Normalized Depth (Z):** Mean depth value of the object relative to the subject

The general form of OIS is given as:

$$OIS = \omega_1 \times S + \omega_2 \times D + \omega_3 \times Z$$

Where:

ω_1 , ω_2, ω_3 are feature weights determined using:

- PCA (variance-based weighting)
- SHAP (model explainability-based importance)
- Ridge Regression or XGBoost (predictive weight learning)

Each object's OIS score reflects its visual and spatial influence within the image.

Scene Similarity Bias (SSB)

As detailed in Section 2.5, the Scene Similarity Bias (SSB) captures how semantically "masculine" or "feminine" the image's scene is, based on CLIP embeddings.

SSB is computed using cosine similarity between the scene embedding $\boldsymbol{E_{image}}$ and two reference vectors:

$\boldsymbol{E_{male}}$ and $\boldsymbol{E_{female}}$.

$$SSB = \cos(E_{image}, E_{male}) - (E_{image}, E_{female})$$

- **Positive SSB**: Scene is more aligned with masculine environments
- **Negative SSB**: Scene is more aligned with feminine environments
- **SSB ≈ 0**: Scene is neutral or ambiguous

This score allows the metric to incorporate **non-object contextual cues**, which often play an overlooked but critical role in biased predictions.

Final UBM Score

The final Unified Bias Metric (UBM) combines both components as a **weighted linear combination**:

$$UBM = \alpha \times OIS + \beta \times SSB$$

Where:

- α and β are scalar weights assigned to the object and scene components respectively
- These can be:
    - **Fixed values** (e.g., $\alpha = \beta = 0.5$)
    - **Optimized via grid search or hyper-parameter tuning**
    - **Learned through regression models like Ridge or XGBoost**
    - **Derived from SHAP-based feature attribution**

This formulation ensures that both localized object-level signals and holistic scene-level context are accounted for in a single, interpretable metric. The UBM can be computed:

- **Per object–person pair**
- **Per image** (aggregated if multiple objects)
- **Per object class** (e.g., average UBM for "handbag")
- **Per dataset** (e.g., global bias score)

The Unified Bias Metric (UBM) offers a novel and flexible approach to diagnosing contextual gender bias in visual datasets. By integrating the spatial influence of gender-associated objects and the semantic alignment of the scene environment, UBM provides a more nuanced, data-driven alternative to traditional bias metrics. Its interpretable, modular design supports extensive experimentation and allows for integration with existing fairness auditing workflows.

2.2.8. Comparative Approaches for UBM Computation

To enhance the robustness, generalizability, and interpretability of the Unified Bias Metric (UBM), this research implements and evaluates nine distinct computational approaches. Each approach combines different weighting strategies ranging from statistical modeling to explainability frameworks to integrate object- and scene-level features into final bias

scores. This analysis offers multiple perspectives on how gender-related contextual bias manifests in visual datasets and enables validation of the stability and fairness of UBM.

Objective of Multi-Approach Design

Since contextual gender bias emerges through both spatial proximity (object-level) and semantic similarity (scene-level), a one-size-fits-all metric may not suffice across datasets or tasks. Therefore, this study adopts a multi-model strategy, where each approach uses a different technique to assign weights to the Object Influence Score (OIS) and Scene Similarity Bias (SSB), including:

- **Unsupervised variance analysis (PCA)**
- **Model-based interpretability (SHAP)**
- **Regularized linear models (Ridge Regression)**
- **Hyperparameter-tuned ensemble methods (XGBoost + Optuna)**

This diversity enables a more nuanced and generalizable interpretation of bias while also improving transparency in how the scores are derived.

Summary of UBM Computational Approaches

Table 2: Summary of UBM Weighting Strategies and Models

Approach	Weighting Strategy	Model Used	Highlights
1	PCA / Random Forest	Regression	Baseline using unsupervised feature variance
2	SHAP Values	Random Forest Classifier	Feature attributions from decision paths
3	SHAP + PCA (Averaged)	RF + PCA	Blends interpretable and unsupervised insights

4	SHAP	XGBoost Classifier	Strong baseline with high explainability
5	SHAP	XGBoost + SMOTE	Handles gender label imbalance using synthetic oversampling
6	SHAP + PCA + Grid Search	XGBoost + SMOTE	Optimizes blend of SHAP–PCA weights via exhaustive search
7	SHAP	XGBoost + Optuna	Hyperparameter-optimized model selection
8	Ridge Regression Coefficients	Ridge Regressor	Linear weighting using coefficient magnitudes
9	SHAP + PCA → Ridge Weighting	SHAP + PCA + Ridge	Final hybrid with explanation and statistical regression fusion

Each model generates UBM scores per image and per object class, which are later aggregated and analyzed by gender category.

Weight Assignment Techniques

Each approach assigns weights to the OIS and SSB components using different logic:

- PCA (Principal Component Analysis): Unsupervised method that allocates weights based on variance in feature space.
- SHAP (SHapley Additive Explanations): Model-agnostic interpretability tool that quantifies feature importance based on predictive contribution.
- Ridge Regression: Regularized linear model that assigns weights based on correlation with target label while controlling overfitting.
- XGBoost + Optuna: Tree-based learning boosted with automated hyper-parameter optimization for fine-tuned scoring.

- Combined Methods: Blend of interpretability (SHAP) and structure (PCA) followed by regression calibration (Ridge).

Evaluation and Visualization Strategy

For each approach, the generated UBM scores are:

- Normalized to [-1, +1] scale
  - Categorized into Male-Biased, Female-Biased, or Neutral using a dynamic threshold
  - Visualized through:
    - Bar plots for object-level bias trends

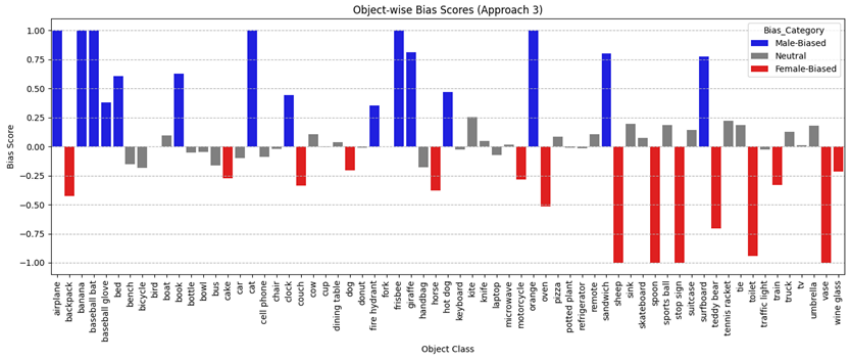

Figure 22: Object-wise Bias Scores computed using Approach 3. Bias categories are color-coded and sorted by object class.

    - KDE (Kernel Density Estimation) plots to show distribution shifts

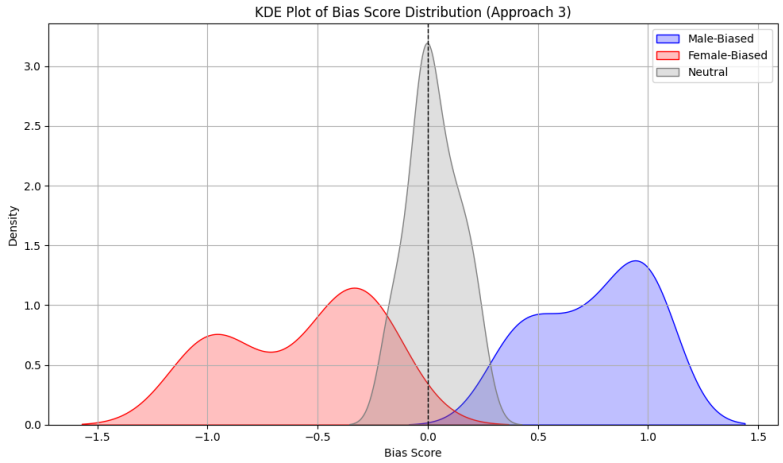

Figure 23: KDE Plot of Bias Score Distribution for Approach 3 (SHAP + PCA Fusion). The distribution illustrates clear clustering of male-biased, female-biased, and neutral object classes.

- SHAP plots to explain feature weightings

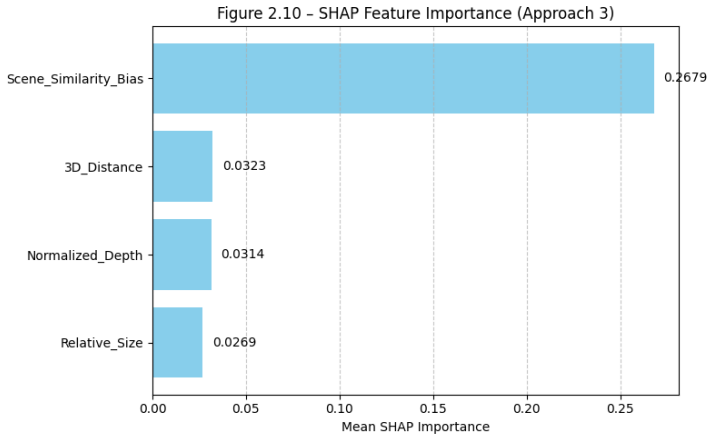

Figure 2.24: SHAP feature importance bar plot indicating contributions of individual features toward bias score prediction

- Statistical Validation: The Mann–Whitney U test is used to compare distributions of bias scores across gender categories, providing non-parametric statistical significance.
- Execution: All nine approaches are fully automated, reproducible, and evaluated across all 852 filtered images in the original dataset.

The comparative framework enables a holistic, multi-angle examination of contextual bias using diverse scoring logic. By triangulating findings across machine learning models,

statistical explanations, and optimization techniques, this section ensures that UBM is not only technically rigorous but also transparent, reproducible, and generalizable. The results of each approach are presented in the following chapter to determine which methods offer the most stable and interpretable bias metrics.

2.2.9. Testing and Implementation

This section outlines the deployment, scalability, and reproducibility of the Unified Bias Metric (UBM) pipeline. To validate its effectiveness, the complete contextual bias detection framework was applied across **eight curated datasets**, each simulating distinct gender-object-scene configurations. The system was designed to test **nine computational approaches**, ensuring broad benchmarking and interpretability. All experiments were executed in **Python 3.10+ using Google Colab**, with support from modular scripts, automated visualization, and structured storage of all outputs.

Dataset Structure and Composition

To rigorously evaluate model behavior under different bias conditions, eight datasets were prepared:

Table 3: Description of Evaluation Datasets Used in the UBM Pipeline

Dataset	Description
Original Dataset	Full set of 852 annotated samples containing person-object-scene triplets.
Male Biased Dataset	Skewed toward male-labeled images (~70–90%) to test sensitivity to male-prevalent environments.
Female Biased Dataset	Contains a higher proportion of female-labeled images (~70–90%).
Balanced Bias Dataset	Equal male and female distribution across all object classes. Used for baseline fairness evaluation.
Neutral Dataset	Composed of objects equally occurring across genders with no prior known bias.
Male Biased (All Genders)	Maintains all object classes but increases male-label frequency, allowing mixed but skewed bias testing.

Female Biased (All Genders)	Same as above but skewed toward female label distribution.
Top Extreme Bias Dataset	Contains only the most gender-polarized object types (e.g., *handbag*, *baseball bat*), stress-testing bias metrics.

Each .csv dataset contains:

Gender, Object_Class, Relative_Size, Normalized_Depth, 3D_Distance, and Scene_Similarity_Bias.

Modular Pipeline Design

The UBM pipeline consists of nine computational approaches, each implemented as a Python class with modular support for:

- Preprocessing: Feature normalization, gender label encoding.
- Model Execution: SHAP value extraction, PCA weighting, Ridge/XGBoost training.
- UBM Scoring: Calculation of α×OIS + β×SSB for every object class.
- Bias Categorization: Bias scores normalized to [−1, +1] and labeled as:
    - Male-Biased (score > threshold)
    - Female-Biased (score < −threshold)
    - Neutral (|score| ≤ threshold)
- Statistical Validation: Mann–Whitney U tests for group comparison.

Runtime per dataset per approach is ~1–2 minutes, and the entire batch of 72 runs (9 × 8) is handled automatically using pipeline_for_9_datasets.py.

Execution Environment and Output Artifacts

- Platform: Google Colab with GPU (when needed)
- Key Libraries: shap, xgboost, optuna, sklearn, matplotlib, seaborn, imbalanced-learn
- Output Directory:

/content/drive/MyDrive/ContextualBias/Analysis/results/<dataset>/<approach>/

For each experiment, the following are saved:

- Bias Score CSVs: Normalized UBM scores + category labels
- SHAP Output Files: Feature attribution summaries
- Visuals: Violin/KDE plots, object-wise bar charts
- Models: Trained classifiers in .pkl format
- Statistical Logs: Mann–Whitney U test summaries

Visual Evaluation Strategy

Visualizations were essential in interpreting model behavior under varying dataset biases. Each dataset-approach combination generated:

- KDE Violin Plots – UBM distribution per object/gender
- SHAP Feature Bar Charts – Feature impact per model
- Bias Category Distributions – Male/Female/Neutral splits
- Statistical Significance – U test output for validation

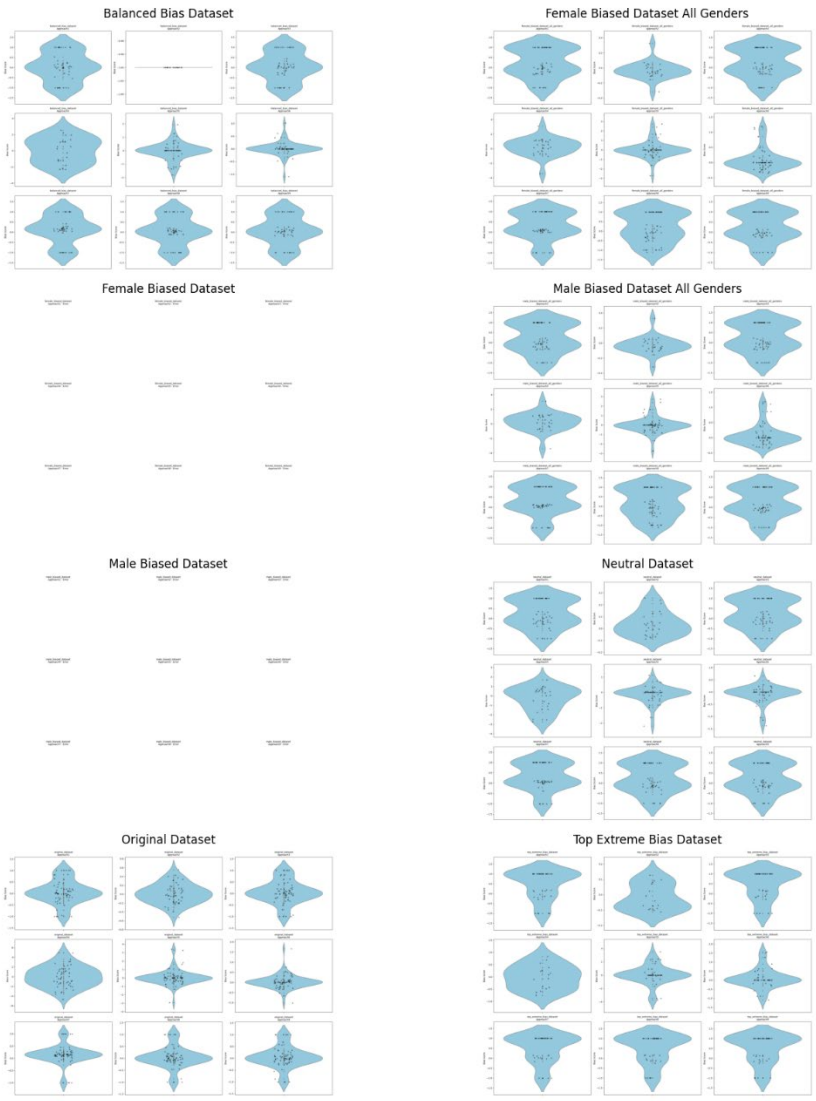

Figure 25: Violin plots visualizing bias score distributions for all nine computational approaches across the eight dataset conditions. The plots reveal clear separation patterns, distribution skewness, and variability in model sensitivity depending on dataset bias (balanced, skewed, or extreme).

This section confirms the scalability, robustness, and transparency of the UBM pipeline. The use of multiple datasets and approaches allows for a nuanced evaluation of contextual gender bias and supports flexible integration with fairness auditing workflows across vision-based AI systems.

2.2.10. Limitations and Assumptions

This section outlines the key limitations and foundational assumptions inherent in the design, implementation, and evaluation of the Unified Bias Metric (UBM) pipeline. While the framework demonstrates strong scalability, explainability, and cross-dataset robustness, several constraints and trade-offs were necessary to balance scope with feasibility. Recognizing these limitations is crucial for properly contextualizing results and guiding future enhancements.

Manual Binary Gender Labeling

UBM relies on manually annotated gender labels for the most prominent person in each image, constrained to a binary classification: **male** or **female**. This binary approach was adopted to maintain consistency with available annotations and to simplify dataset curation. However, it inherently excludes non-binary, gender-fluid, or ambiguous identities, limiting the framework's inclusivity and generalizability beyond the binary paradigm.

Bias in Pre-Trained Models

The pipeline employs several state-of-the-art pre-trained models **YOLOv8** (object detection), **CLIP** (scene embedding), and **Places365** (scene classification) which were trained on large-scale datasets that may encode latent societal biases. Although these models significantly improve automation and accuracy, their outputs may inadvertently propagate existing biases into UBM’s feature space. The pipeline currently assumes model neutrality and does not explicitly correct for potential upstream bias.

Dataset Imbalance and Representation

Despite using curated datasets such as the Balanced and Neutral sets, residual object-gender imbalances persist due to real-world data distributions. Some object classes may naturally appear more frequently with one gender, introducing structural bias into the dataset. The pipeline assumes these distributions are representative unless deliberately corrected, meaning UBM scores may reflect both natural and dataset-induced bias.

Fixed Scene and Object Semantics

Scene and object semantics are extracted based on fixed model-generated labels (e.g., "handbag," "kitchen"), which are treated as universally meaningful. However, these labels can carry cultural, regional, or contextual variations in meaning. Relying on static definitions may oversimplify complex social interpretations and reduce metric sensitivity to contextual nuance.

Threshold-Based Categorization

UBM scores are categorized into **Male-Biased**, **Female-Biased**, or **Neutral** using a threshold-based classification scheme. While effective for summarization, this method may **oversimplify subtle variations** in contextual influence and may not capture gradient shifts or cumulative bias effects. Fine-grained differences could be lost in rigid classification, especially near threshold boundaries.

The UBM pipeline makes several controlled assumptions regarding gender classification, pre-trained model reliability, and data structure to ensure consistency and computational efficiency. While these assumptions enable a robust and interpretable analysis pipeline, they also introduce boundaries in inclusivity, flexibility, and generalizability. Acknowledging these limitations is essential for contextualizing the findings and motivating future work to advance fairness, intersectionality, and social relevance in AI-driven bias detection.

2.2.11. Commercialization Aspects of the Metric

The Unified Bias Metric (UBM) developed in this research presents strong commercialization potential as an interpretable, scalable, and modular solution for auditing contextual gender bias in image datasets. Designed with extensibility and automation in mind, UBM can be deployed in diverse real-world contexts including dataset auditing, AI fairness validation, compliance reporting, and academic benchmarking. By capturing both object-level spatial dynamics and scene-level semantic

cues, UBM provides a uniquely holistic view of visual bias going beyond traditional frequency or co-occurrence methods.

Its modular implementation, explainability via SHAP, and compatibility with popular ML libraries position UBM as a candidate for integration into MLOps pipelines, AutoML platforms, or cloud-based dashboards. Whether packaged as a Python library, CI/CD plugin, or interactive web tool, UBM offers value to developers, researchers, and policy auditors alike. Its visual outputs (e.g., KDE plots, SHAP charts) further support usability across technical and non-technical stakeholders, making it a practical and versatile tool for promoting algorithmic fairness in computer vision.

2.3. Audio Bias Score: Bias Detection Metric for Gender Bias Detection in Audio Datasets.

2.3.1. Selection Of Features for Equation Building.

The selection of features which were focused on when building the equation were done through studying previous research based on audio-based gender classification systems by analysing which features were considered when building such systems. Through the research studies the primary features considered were found to be pitch [31], Amplitude [32], Energy [33], Formants, Intonations and Mel-Frequency Cepstral Coefficients (MFCCs) [34] [35] [36] [37].

Intonations [38] [39] and formants [40] are dependent on the language and vary among cultures and ethnic groups of the speaker as the metric is built independent of the language and race with the sole focus on identifying Gender bias in datasets these features were not considered when building the metrics. Also, it is stated by Bailey Et al. (2021) [41] that models based on raw audio are more robust to gender biased than ones based on hand-crafted features, such as mel-spectrograms [41]. Therefore, raw audio-based features were only focused on when building the metric.

Additionally, Number of Audios and voice activity per gender is used in the metrics. Where number of audios are considered as a way for identifying class imbalances and Voice activity to quantify even if the classes among genders are properly balanced with the difference of voice activity it can affect the performance of a model trained on such a dataset.

Gender based variations naturally exist in the factors pitch, Amplitude and Energy levels, where pitch is inherently higher for female speakers. Energy levels and amplitude also display consistent high or low values for a specific gender depending on the speech characteristic. For the quantified bias to not be affected by these inherent behaviours the standard deviations of the pitch, amplitude and energy levels were used.

In conclusion, the count of audio files per gender, Voice activity per gender, Standard deviation of Amplitudes, Energy and pitch per gender was used to build the equation.

Selected features explained.

I. Pitch

Pitch is one of the most fundamental characteristics of human speech and is directly associated with the fundamental frequency produced by the vibration of the vocal cords. In audio signals, pitch is perceived as the frequency of a sound wave and is typically measured in Hertz(Hz). It serves as a powerful and widely used indicator for gender classification systems, this is mainly due to the physiological differences between male and female speakers. Male speaker generally possess thicker and longer vocal folds which vibrate at a lower frequency range which is approximated to fall between the range of 85Hz to 180Hz . In contrast, female speakers tend to have a shorter and thinner vocal fold that vibrate at a higher frequency.

However, pitch is not solely related to the speakers' gender, it is also modulated by the emotional state, linguistic context and individual speech habits. These factors introduce natural variation within gender groups which, if not accounted for, could misinterpret pitch based measurements .

To avoid conflating natural biological difference with potential systematic bias, this study utilizes the standard deviation of pitch with each gender group rather than sum or the mean of the pitch values. This approach enable the identification of inconsistencies in pitch distribution across genders that are reflective of datasets construction bias rater than innate vocal characteristics.

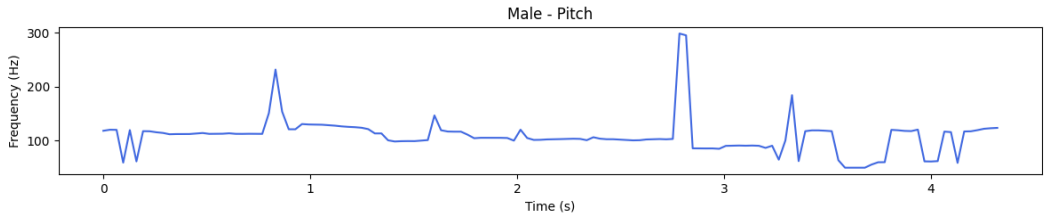

Figure 26 Single male speaker pitch distribution over time.

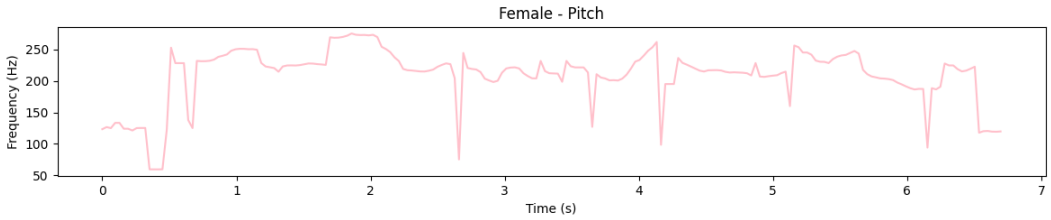

Figure 27 Single female speaker pitch distribution over time.

II. Amplitude.

Amplitude reflects the strength or magnitude of and audio signal, corresponding to the level of air pressure fluctuations produced during speech. In waveform representation, amplitude is depicted by the height of the wave and is closely related to the perceived loudness or intensity of speech. While, higher amplitude values generally denote louder speech, the amplitude of an audio signal can be influenced by several factors, including speaking style, vocal effort and environmental conditions.

Physiologically, amplitude differences between genders may occur due to varying subglottal pressures and vocal fold dynamics. However, amplitude is also susceptible to systematic factors such as inconsistencies in microphone placement, room acoustics and speaker posture. These factors can disproportionately affect one gender over another, especially in datasets compiled from diverse sources or uncontrolled recording environments.

In order to address these concerns the study uses the standard deviation of amplitude per gender, rather than relying on absolute amplitude values. This statistical measure provides insights into the internal variability of amplitude within each gender, which offers a assessment of how loudly or softly speakers are represented in the dataset in specific gender.

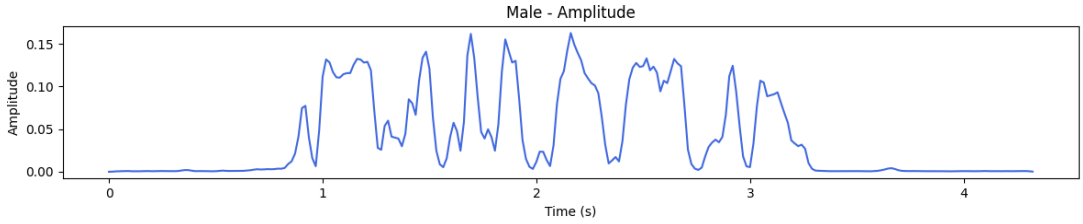

Figure 28 Single male speaker amplitude distribution over time.

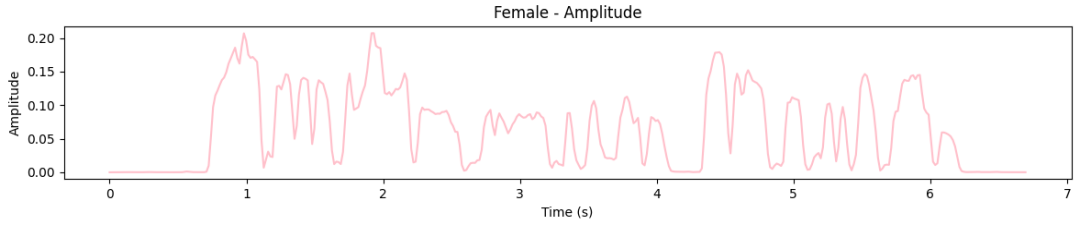

Figure 29 Single female speaker amplitude distribution over time.

III. Energy.

Energy in an audio signal is a measure of total power of the sound wave over time. It is typically calculated by squaring the amplitude values of a signal within a given ti,e window and summing them to obtain cumulative acoustic power. This feature captures the force or effort behind speech production and plays a significant role in applications such as speaker activity detection, emotion analysis and speech clarity assessments.

The energy level may vary due to lung capacity, vocal effort and articulation habits. However, similar to amplitude energy is influenced by physiological as well as technical and environmental variables. Recording inconsistencies microphone sensitivity etc. If one

gender is disproportionately affected by such external variables the dataset may exhibit bias.

Similar to pitch and amplitude in order to avoid the influence of these confounding factors, standard deviation of energy is calculated separately for each gender in this study. Using standard deviation allow to evaluate how uniformly energy is distributed within a gender.

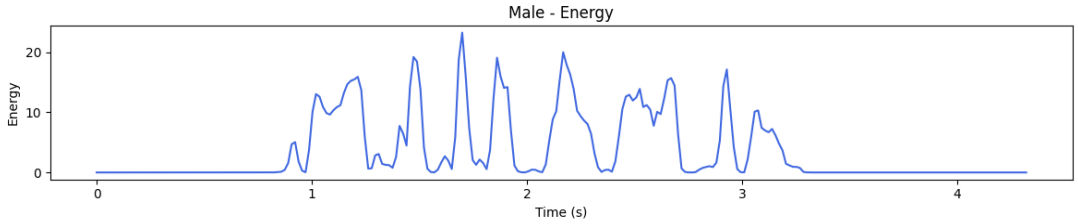

Figure 30 Single male speaker energy distribution over time.

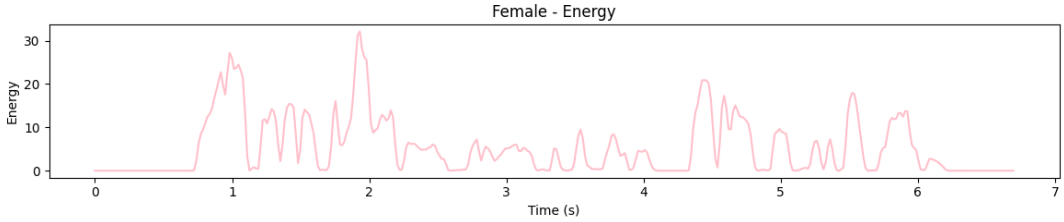

Figure 31 Single female speaker energy distribution over time.

IV. Voice activity per gender.

Voice activity is defined as the temporal duration during which a speaker is actively producing speech excluding periods of silence, background noise or non-verbal sounds. This metric is crucial for assessing not just the presence of the audio in samples in a dataset but the actual contribution of each gender in terms of usable content-rich speech.

In many audios datasets equality in the number of audio files per gender does not necessarily equate to equality in representation. While male and female speaker maybe present in equal numbers, if one gender consistently speaks longer durations per audio clip the datasets will inadvertently favor that group in terms of speech content. This

discrepancy can influence model performance by providing more training data and phonetic diversity for one gender leading to biased learning outcomes.

To account for this voice activity per gender is computed to quantify the actual spoken duration across all samples within each gender category. By incorporating this metric, the study ensures that bias detection is not limited to numerical class balance but extends to qualitative aspects of speech as well.

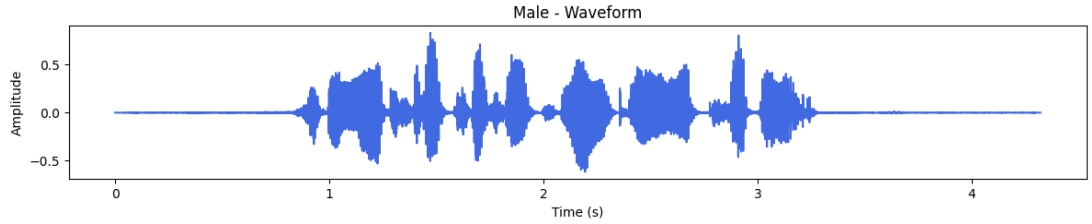

Figure 32 Single male speaker waveform

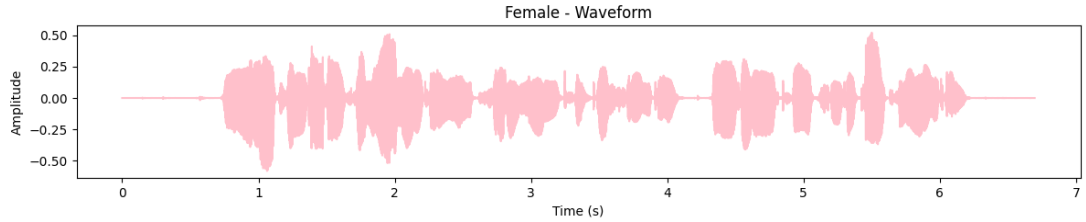

Figure 33 Single female speaker waveform

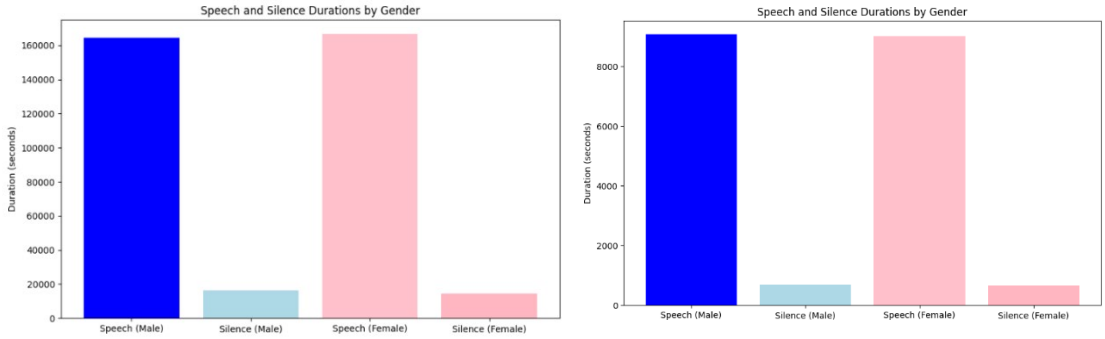

Figure 34 Voice Activity of Genders in LibriSpeech 360 (Left) and LibriSpeech 100 (Right)

V. Number of audio samples per gender.

The count of audio samples per gender is a fundamental measure of identifying class imbalance in speech datasets. And equal number of recordings between male and female speaker can directly affect the performance of the machine learning models training on such data. Class imbalance often leads to biased model predictions favoring the overrepresented class while under performing on the minority class. This class imbalance not only affects classification accuracy but also fairness in real-world deployment.

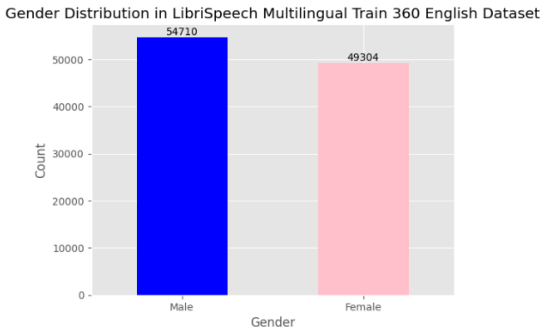

Figure 35 Gender Distribution LibriSpeech Multilingual Train 360

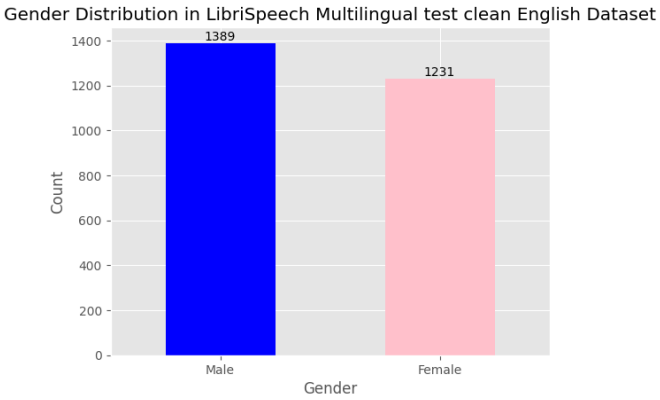

Figure 36 Gender Distribution LibriSpeech Multilingual Test dataset

2.3.2. Building the dataset.

Introducing the datasets.

I. Common Voice Dataset.

The Mozilla Common Voice dataset [42] is a large-scale, open-source collection of voice recordings contributed by volunteers around the world. It was created to help train machine learning models for automatic speech recognition (ASR) and other speech-related tasks. As of its latest versions, Common Voice includes millions of samples across more than 90 languages and dialects, making it one of the most linguistically diverse open-source speech corpora available.

Each audio sample is paired with its corresponding transcription, and the dataset is structured with detailed metadata including speaker gender, age, accent, and recording device information. This rich set of metadata makes Common Voice especially useful in bias analysis and fairness evaluation across demographic variables. Due to the crowd-sourced nature of data collection, it exhibits natural variations in recording quality, pronunciation, pitch, and speech speed, which are valuable in simulating real-world conditions. Its diverse linguistic representation and demographic labeling make it an ideal candidate for building a base dataset for gender bias evaluation in audio data.

II. LibriSpeech Dataset.

The LibriSpeech dataset [43] is one of the most well-known corpora in the field of speech processing. It is derived from audiobooks that are part of the public domain LibriVox project. The recordings are primarily in English and are accompanied by time-aligned text transcriptions. The dataset is divided into various subsets based on quality and speaker characteristics, such as "train-clean-100", "train-clean-360", and "train-other-500".

LibriSpeech offers high-quality, 16kHz recordings and includes over 1,000 hours of speech. Its structure provides speaker labels and gender information, making it suitable for demographic analysis. Because the speakers are reading text in a controlled manner, the recordings are generally of high clarity and consistent linguistic content, which is advantageous for extracting clean acoustic features such as pitch, amplitude, and energy. It has been widely used as a benchmark dataset in both traditional and deep learning-based speech recognition research.

III. LibriSpeech Multilingual dataset.

The LibriSpeech Multilingual (LibriVox Multilingual Speech) [44] dataset is an extension of the original LibriSpeech, aimed at promoting research in multilingual ASR systems. It contains recordings from the LibriVox project in multiple languages including French, German, Spanish, and Italian, among others. Just like its predecessor, it provides audiobooks read by volunteers, but this time in several different languages, each with aligned text transcriptions.

This dataset is essential for studying language-independent speech characteristics. It allows researchers to investigate how acoustic features vary across languages and whether bias detection models trained on language-agnostic features remain effective. Because it includes gender and speaker ID metadata, it supports comparative gender bias studies across languages. It's particularly useful for developing and validating language-neutral metrics, as intended in this research.

IV. TED-LIUM Dataset.

The TED-LIUM dataset [45] is constructed from TED talks, which are presentations delivered by speakers from around the world on a broad range of topics. The dataset includes transcriptions aligned with audio at the word level. It features diverse speakers with varying accents, intonations, speech styles, and language proficiencies,

offering a realistic and challenging set of conditions for speech recognition and bias analysis.

What sets TED-LIUM apart is the variability in speech caused by spontaneous delivery, public speaking nuances, and emotional expression. Unlike read speech in LibriSpeech, TED-LIUM features natural, conversational language. This introduces fluctuations in pitch, energy, and speaking rate—core components used in your bias metric. The dataset includes speaker metadata such as gender and speaker identity, allowing researchers to evaluate how models generalize across demographic and expressive variances.

V. AMI meeting corpus.

The AMI (Augmented Multi-party Interaction) Meeting Corpus [46]is a dataset that captures real-world meeting conversations among multiple participants. It consists of approximately 100 hours of meeting recordings, including both audio and video, transcriptions, dialogue acts, and metadata on speaker roles and demographics.

The audio in AMI is characterized by overlapping speech, background noise, and informal conversational dynamics, reflecting highly realistic communication scenarios. The corpus was recorded using multiple microphone types (individual headset mics and room mics), introducing variations in audio quality. This diversity in acoustic environments and speaker interactions makes it a valuable resource for testing the robustness of audio features such as energy distribution and voice activity under more complex, real-world conditions.

The availability of speaker role, gender, and conversation style information also makes AMI ideal for testing how gender bias manifests not just in solo speech but in interactive settings. This dataset is particularly useful for examining how bias scores may be influenced by group dynamics or speech interruptions, adding depth to the validation process of the proposed metric.

Combination of Each of these datasets' different splits as well as subsets for different languages were used to build one training and testing dataset.

	A	B	C	D	E	F	G	H	I	J	K
1	Dataset split	Count_male	Count_female	voice_activity_male	voice_activity_female	energy_male	energy_frmale	amplitude_male	amplitude_female	pitch_male	pitch_female
2	Librispeech Englsih Dataset										
3	Dev-clean(337MB)	1374	1329	8946.75	9020.14	1282.84	1197.81	450.94	708.26	42.25	45.56
4	Dev-other(314MB)	1450	1414	8468.01	8319.99	1289.06	1203.7	387.43	409.11	38.62	44.26
5	test-clean(346MB)	1389	1231	9073.79	9013.66	1188.67	1119.96	376.01	376.68	44.34	55.15
6	test-other 328	1561	1378	8651.95	8708.53	1298.58	1067.23	511.5	335.28	46.32	46.7
7	train clean 100 6.3GB	14342	14197	164431	166619	1189.73	1116.82	616.51	432.79	46.08	47.62
8	train clean 360 23GB	54710	49304	619310	574195	1146.69	1188.46	528.36	655.45	46.84	48.15
9	Multiligual Librispeech Dataset										
10	Italian_train	32609	27014	434400	362925	1341.03	1022.05	984.58	549.15	38.38	47.36
11	Italian_test	694	568	9207.42	7893.26	1040.45	1012.51	643.98	357.28	35.54	34.33
12	Italian_dev	544	704	7596.22	9516.49	1365.35	954.68	772.14	403.68	33.93	38.68
13	portuguese_train	12865	24668	166937.9	327847.42	988.97	790.62	334.57	285.12	40.22	52.17
14	portuguese_test	447	424	6201.42	5975.38	939.32	833.45	295.7	238.28	34.46	40.03
15	portuguese_dev	508	318	7405.32	4677.85	948.29	786.19	275.72	306.25	58.76	57.08
16	polish_train	18691	6352	262854.4	91507.09	844.49	851.29	242.28	320.39	45.25	63.27
17	polish_test	255	265	3479.44	3556.55	744.2	1110.75	216.71	386.9	42.02	47.08
18	polish_dev	229	283	3175.68	3750.11	1135.5	767.06	290.79	190.69	46.75	39.22
19	Common voice Dataset										
20	arabic_train	3612	4624	13299.15	16075.8	2259.04	2636.67	744.12	1166.16	22.54	19.64
21	br_test	776	209	1906.84	517.66	1905.71	1785.44	1146.51	1041.81	20.47	18.21
22	br_validation	862	157	1981.46	407.88	1722.29	1951.06	937.62	686.23	21.97	27.73
23	cnh_train	192	163	567.32	435.25	1356.7	1257.75	418.67	665.36	25.88	11.35
24	cnh_other	1176	815	2646.93	2583.07	624.48	1451.89	848.41	969.38	17.75	15.37
25	cnh_validated	608	467	2045.78	1421.24	1300.81	1385.5	849.2	911.84	19.29	15.25
26	cv_train	1046	105	4177.5	427.36	1129.06	2418.14	1340.83	478.76	25.14	22.44
27	cv_test	324	105	1396.79	449.87	2340.06	3595.41	1192.15	907.62	23.82	18.65
28	cv_validated	9366	4816	39611.78	19852.82	1731.49	2754.49	1087.17	733.02	20.7	21.76
29	cy_train	3510	2659	14884.86	11492.82	1978.5	2238.04	736.05	963.01	22.24	24.68
30	cy_validation	1538	1246	6689.72	5683.38	2176.02	1853.45	827.53	944.25	23.6	23.98
31	cy_test	693	595	3043.32	2587.22	1868.41	1678.06	942.1	970.84	24.35	29.38
32	cy_other	6215	3789	26737.4	16813.16	1834.56	1653.12	919.3	920.27	23.36	31.34
33	cy_validated	36205	27229	141901.4	112437.26	1935.57	1993.45	1004.07	1052.17	21.67	23.62

Figure 37 Extracted Features stored in CSV.

Processing and building the Dataset.

Unlike classification tasks that have clearly defined labels (e.g., male vs. female), bias detection lacks an inherent target variable. There is no predefined value that indicates how biased a sample or dataset is unless an external benchmark or metric is imposed. Since the features considered in the proposed metric, such as standard deviations of acoustic attributes are relatively novel in this context, no existing benchmark bias scores were available for supervised training or evaluation. To address this the creation of a synthetic bias score needed to be done.

In order to create a synthetic bias score a controlled data augmentation needed to be performed on the existing extracted values. The idea of the controlled data augmentation approach was to simulate bias by systematically altering feature values associated with one gender group while keeping the other unchanged, and vice versa. This allowed for the

controlled introduction of disparities between genders along the selected features, thereby creating a gradient of bias that could be quantitatively tracked.

The augmentation process involved the following steps:

- The curated dataset was duplicated, resulting in two equal-sized subsets—each containing the same audio files and corresponding feature vectors.
- In the first half of the augmented dataset, feature values for male speakers were kept unchanged, while feature values for female speakers were reduced in controlled increments of 5%. This reduction was applied to all core features. For example, a feature value of 0.20 for a female speaker would be modified to 0.19 (5% reduction), 0.18 (10% reduction), and so forth across different versions of the sample.
- In the second half of the dataset, the process was reversed: female speaker features were kept constant, while male speaker features were incrementally reduced in the same 5% steps.
- Each modification introduced a known level of imbalance between the two gender groups, allowing a bias score to be associated with each sample. This score reflected the degree of artificial bias introduced via feature manipulation. For instance, an unmodified sample was assigned a score of 0.0, while a sample with

a 5% reduction in one gender's features was assigned a score of 0.05, with the scale extending to higher values such as 0.10, 0.15, and so on.

- The final augmented dataset thus contained samples with associated bias scores ranging from 0.0 to 1.0, representing a linear and interpretable scale of bias intensity introduced through controlled feature perturbations.

	A	B	C	D	E	F	G	H	I	J	K
1	count_mal	count_fem	voice_activ	voice_activ	energy_ma	energy_fen	amplitude_	amplitude_	pitch_male	pitch_fema	score
2	693	34.65	50.722	2.5361	1868.41	93.4205	942.1	47.105	24.35	1.2175	4.75
3	6215	310.75	445.6233	22.28117	1834.56	91.728	919.3	45.965	23.36	1.168	4.75
4	36205	1810.25	2365.023	118.2512	1935.57	96.7785	1004.07	50.2035	21.67	1.0835	4.75
5	71487	3574.35	3574.59	178.7295	2100.74	105.037	983.69	49.1845	25.58	1.279	4.75
6	24650	1232.5	1363.977	68.19883	1942.62	97.131	910.38	45.519	26.62	1.331	4.75
7	6237	311.85	358.2067	17.91033	1330.55	66.5275	808	40.4	28.55	1.4275	4.75
8	8448	422.4	623.6887	31.18443	1653.63	82.6815	941.48	47.074	25.2	1.26	4.75
9	4455	222.75	291.18	14.559	1640.67	82.0335	1181.14	59.057	26.2	1.31	4.75
10	12505	625.25	691.9583	34.59792	1827.01	91.3505	994.97	49.7485	21.45	1.0725	4.75
11	4304	215.2	239.1013	11.95507	1450.97	72.5485	834.85	41.7425	29.93	1.4965	4.75
12	13006	650.3	1282.221	64.11103	1176.53	58.8265	1031.47	51.5735	23.47	1.1735	4.75
13	4643	464.3	304.4187	30.44187	1734.59	173.459	789.55	78.955	26.61	2.661	4.5
14	14814	1481.4	909.0247	90.90247	1598.69	159.869	756.15	75.615	28.78	2.878	4.5
15	11029	1102.9	735.8437	73.58437	1797.3	179.73	924.51	92.451	28.4	2.84	4.5
16	11441	1144.1	954.402	95.4402	1708.79	170.879	955.86	95.586	35.1	3.51	4.5
17	15555	1555.5	1052.119	105.2119	1748.45	174.845	947.13	94.713	28.01	2.801	4.5
18	6708	670.8	53.704	5.3704	2031.87	203.187	1471.13	147.113	23.77	2.377	4.5
19	20530	2053	289.991	28.9991	1465.55	146.555	935.74	93.574	28.41	2.841	4.5
20	15500	1550	1060.504	106.0504	1409.54	140.954	892.18	89.218	28.14	2.814	4.5
21	5814	581.4	420.549	42.0549	1386.29	138.629	870.94	87.094	28.73	2.873	4.5
22	12373	1237.3	125.738	12.5738	2029.45	202.945	7544.28	754.428	21.88	2.188	4.5
23	13525	1352.5	760.981	76.0981	1482.32	148.232	1080.08	108.008	26.07	2.607	4.5
24	22418	3362.7	1239.951	185.9927	1834.95	275.2425	1104.5	165.675	45.62	6.843	4.25
25	14418	2162.7	936.6322	140.4948	1504.41	225.6615	1115.3	167.295	35.76	5.364	4.25
26	2781	417.15	145.901	21.88515	1831.28	274.692	1050.33	157.5495	47.27	7.0905	4.25
27	15720	2358	988.0373	148.2056	1957.85	293.6775	1163.17	174.4755	23.38	3.507	4.25

Figure 38 Dataset with controlled Data Augmentation Technique Applied.

2.3.3. Building And Training the Equation.

The training dataset was created by applying the controlled data augmentation technique to base dataset. As the intended method of developing a predictive equation was linear regression, it was compulsory to assess whether the dataset met the key assumptions required for its application. Therefore, the dataset was evaluated on the assumption's linearity, independence of observations, Homoscedasticity, normality of residuals and absence of multicollinearity.

Explaining the assumptions of Linear regression.

I. Linearity.

Linearity refers to the assumption that the relationship between the dependent variable (Y) and the independent variables (X_1, X_2, ..., X_n) is linear in parameters. That is, changes in the input features lead to proportional changes in the output. The standard linear regression model is represented by the equation

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \varepsilon$$

- Y: Dependent variable (bias score)
- $X_1, X_2, \ldots, X_n$: Independent variables (e.g., standard deviations of audio features)
- β_0: Intercept
- $\beta_1, \ldots, \beta_n$: Coefficients of the predictors
- ε: Error term (residual)

To verify this assumption, scatter plots were initially used to visually inspect the relationships between each predictor and the response variable. Additionally, Pearson's correlation coefficient and the correlation matrix were utilized as quantitative measures to assess the strength and direction of linear associations between pairs of variables. These tools collectively ensured that the linearity condition was sufficiently met.

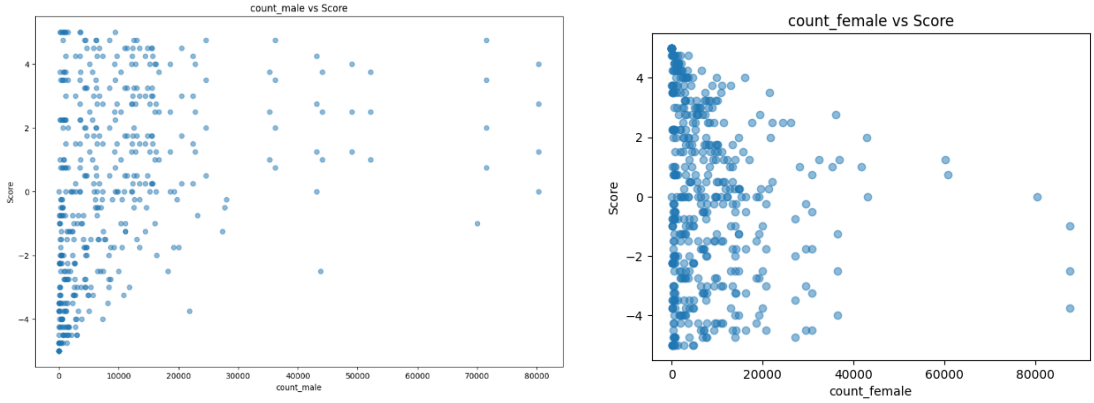

Figure 39 Linearity check : Count Vs Score - Male count(Left), Female Count (Right)

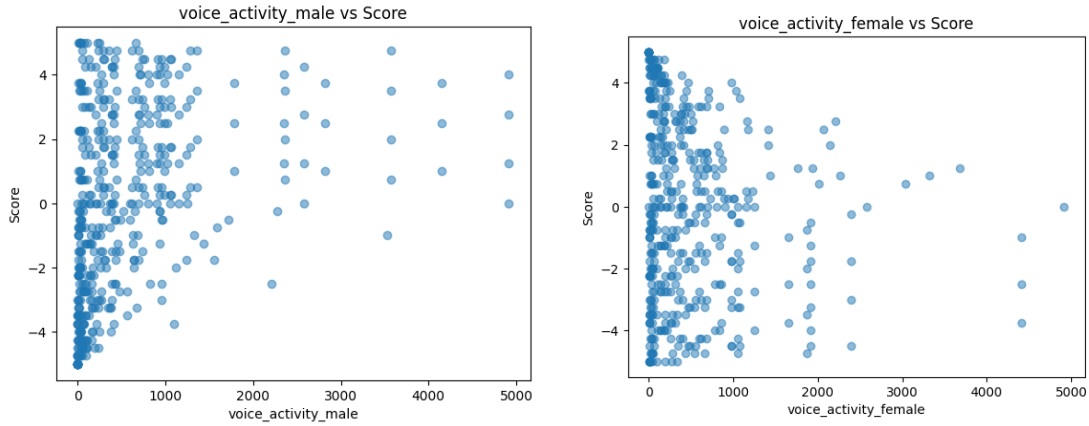

Figure 40 Linearity check : Voice Activity Vs Score - Male Voice Activity(Left), Female Voice Activity (Right)

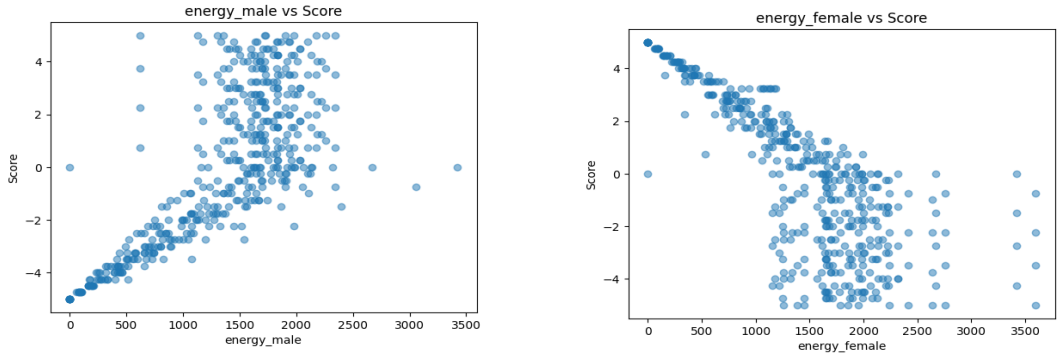

Figure 41 Linearity check : Std. Energy Vs Score - Male Std. Energy(Left), Female Std. Energy (Right)

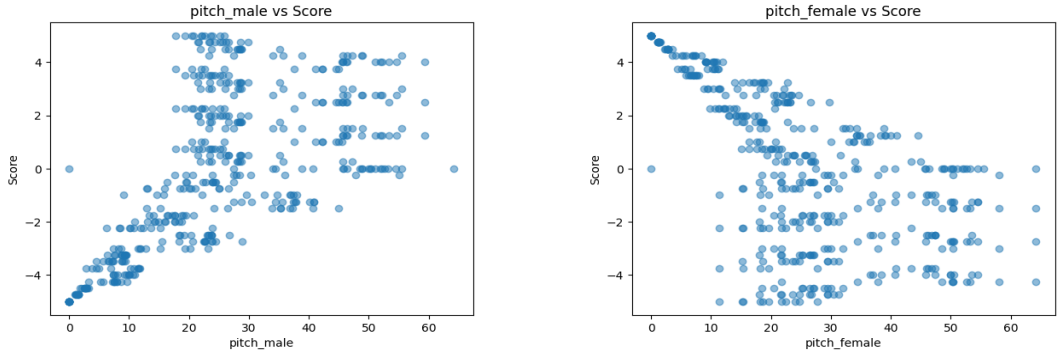

Figure 42 Linearity check : Std. Pitch Vs Score - Male Std. Pitch(Left), Female Std. Pitch (Right)

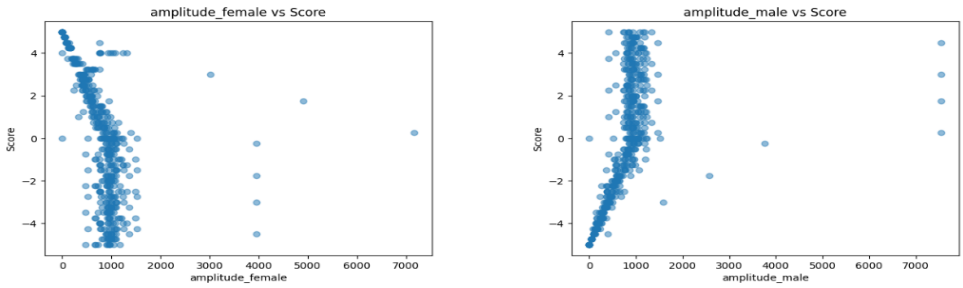

Figure 43 Linearity check : Std. Amplitude Vs Score - Male Std. Amplitude(Left), Female Std. Amplitude (Right)

Pearson Correlation with Score:

	Feature	Pearson Correlation
4	energy_male	0.753126
8	pitch_male	0.628183
6	amplitude_male	0.426615
0	count_male	0.342894
2	voice_activity_male	0.332102
1	count_female	-0.131895
3	voice_activity_female	-0.138098
7	amplitude_female	-0.445231
9	pitch_female	-0.604615
5	energy_female	-0.785568

Figure 44 Pearson's Correlation Score

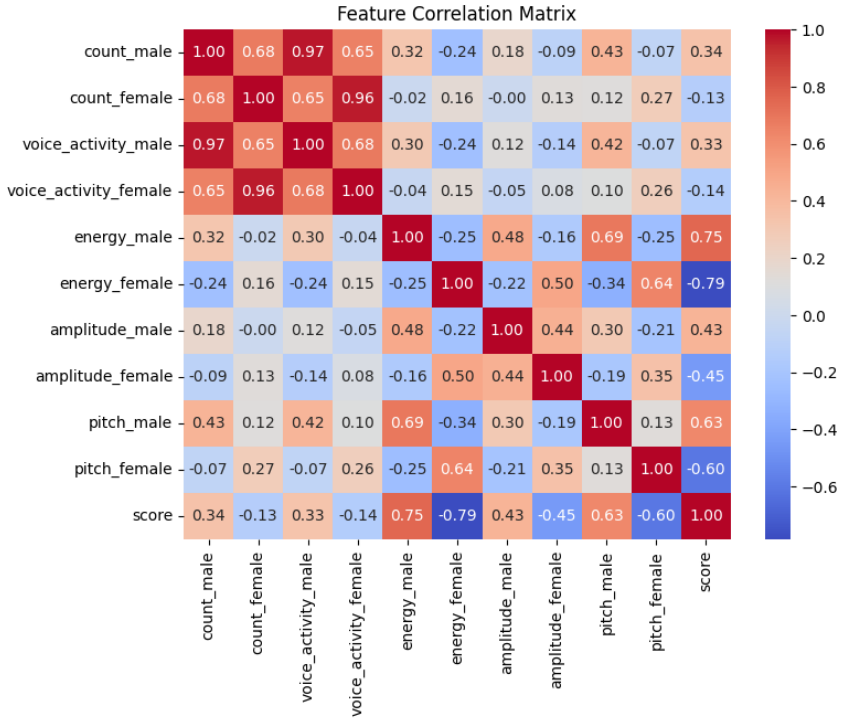

Figure 45 Correlation matrix

II. Independence of Observations.

This assumption states that each observation in the dataset should be independent of the others, meaning that the residuals (errors) of the model should not be correlated across observations. This is particularly critical in time series or sequential data, where autocorrelation may exist. The violation of this assumption can result in underestimated standard errors, leading to overconfident inferences.

To assess this, the Durbin-Watson statistic was computed, defined as:

$$DW = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2}$$

Where:

- e_t: residual at time *t*,

Values of DW close to 2 indicate no autocorrelation, whereas values below 1 or above 3 suggest positive or negative autocorrelation respectively. The datasets' Durbin-Watson statistic was equal to the value of 1.0276 displaying a possible positive autocorrelation.

III. Homoscedasticity.

Homoscedasticity refers to the requirement that the residuals exhibit constant variance across all levels of the independent variables. When this condition is violated the standard errors of the regression coefficients become unreliable, which affects hypothesis testing. This assumption was evaluated by plotting the residuals versus the fitted values. A random scatter of points around the horizontal axis with no discernible pattern is indicative of homoscedasticity. Furthermore, formal statistical tests such as the Breusch-Pagan test and White's test was employed to detect the presence of non-constant variance in the residuals. In this study, the visual inspection of residual plots did not reveal any funnel-shaped patterns, thereby confirming the homoscedasticity assumption.

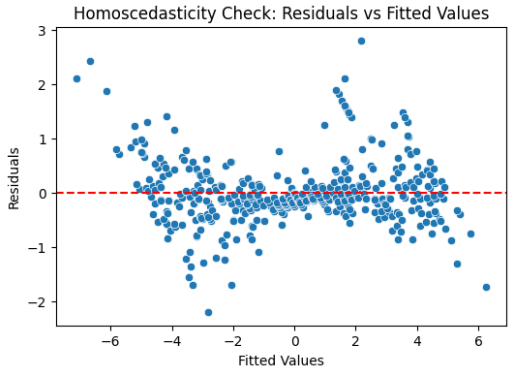

Figure 46 Homoscedasticity check

Apart from the visual inspections the statistical tests Breusch-Pagan test and White's test displayed the following values respectively

- Breusch-Pagan test
    - LM stat: 82.087
    - LM p-value: 1.95×10^{-13}
    - F-stat: 9.741
    - F p-value: 8.54×10^{-15}

  Given that the p-values for both the LM statistic and F-statistic are significantly less than the conventional threshold of 0.05,the null hypothesis was rejected suggesting that heteroscedasticity is present.
- White's test
    - White's test statistic: 210.34
    - Degrees of freedom: 88
    - p-value: 5.12×10^{-12}

  Given the extremely low p-value, the null hypothesis was rejected, concluding that heteroscedasticity is present in the model residuals.

IV. Normality of Residuals.

The normality assumption stipulates that the residuals of the regression model should be normally distributed. This assumption is essential to ensure the validity of confidence intervals and hypothesis tests concerning the regression coefficients. To evaluate this, a Q-Q (Quantile-Quantile) plot was generated, which plots the quantiles of the residuals against the theoretical quantiles of a standard normal distribution. If the residuals are normally distributed, the points on the Q-Q plot should align closely with the 45-degree diagonal line.

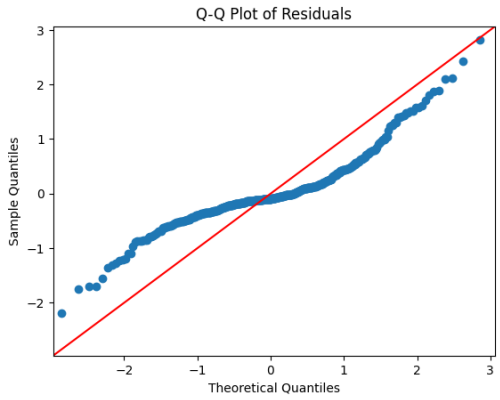

Figure 47 Q-Q Plot of residuals

In addition to the Q-Q plot, the Shapiro-Wilk test and the Kolmogorov-Smirnov test was used for statistical validation of normality and they displayed the results as follows:

- Shapiro-Wilk Test
    - Test Statistic: 0.957
    - P-value: 2.41×10^{-10}

    Given that the p-value is significantly smaller than the conventional threshold of 0.05, we reject the null hypothesis (H_0), indicating that the data is not normally distributed.

- Kolmogorov-Smirnov Test.
    - Test Statistic: 0.293
    - P-value: 1.33×10^{-35}

    Given that the p-value is extremely small (much smaller than the threshold of 0.05), we reject the null hypothesis (H_0), concluding that the data does not follow a normal distribution.

V. Absence of Multicollinearity.

Multicollinearity occurs when two or more independent variables are highly correlated with each other, leading to inflated standard errors and unreliable estimates of regression coefficients. The Variance Inflation Factor (VIF) was employed to diagnose multicollinearity among the predictors, calculated as:

$$VIF_i = \frac{1}{1 - R_I^2}$$

Where:

- R_I^2: coefficient of determination.

Variance Inflation Factor (VIF):

	Feature	VIF
0	const	13.486425
1	count_male	41.965076
2	count_female	35.075759
3	voice_activity_male	42.444733
4	voice_activity_female	35.483454
5	energy_male	4.354491
6	energy_female	4.304774
7	amplitude_male	3.281165
8	amplitude_female	3.228986
9	pitch_male	4.993634
10	pitch_female	4.089872

Figure 48 Variance Inflation Factor

The dataset violated Normality, Multicollinearity, Homoscedasticity, and Independence of the observations. Therefore, simple linear regression could not be used for equation building. To build the equation in a way where these violations will not affect the performance of the metric the choice of methods Symbolic regression and Polynomial

regression with L2 (Ridge) regularization were available. Each of these methods were used to build an equation after which the performance was compared, and a final decision was made based on the performance of each equation.

Building the equation: Method - Symbolic regression.

The training dataset was utilized to develop a predictive model using symbolic regression, an evolutionary algorithm-based technique that searches the space of mathematical expressions to find an optimal equation that best fits the data. To optimize the model's performance, a hyperparameter tuning process was carried out by experimenting with different configurations. The hyperparameters explored included:

- Population size: [500, 1000, 1500]
- Number of generations: [5, 10, 15]
- Crossover probability (p_crossover): [0.6, 0.7]
- Point mutation probability (p_point_mutation): [0.05, 0.1]

Following extensive experimentation, the best-performing configuration was found to be:

- Population size: 500
- Point mutation probability: 0.05
- Crossover probability: 0.7
- Generations: 15

This combination produced the lowest Mean Squared Error (MSE) value of 6.64160, indicating good predictive accuracy. The symbolic regression algorithm subsequently generated a closed-form mathematical expression that models the relationship between the selected input features and the target variable as follows:

$$A = (0.460 - x_6) * ((x_7 - x_6) * (x_9 * x_7))$$

$$B = ((((-0.015 - x_4) * (-0.776 + X_5)) - \frac{X_6}{X_8}) + X_6$$

$$C = (-0.015 - X_4) + \left(\frac{X_6}{(X_1 + X_5)}\right) + 0.959 - \left(\frac{X_5}{(X_{8*}(X_4 + X_5))}\right) - ((x_9 + x_6) + \left(\frac{x_1 + (x_9 - x_3)}{X_2} * X_7\right)) + \left(\frac{X_1 * (X_0 - 0.781)}{X_8}\right) + (\frac{-0.752}{-0.055})$$

$$D = \frac{x_8}{x_4}$$

$$E = x_9 + (((-0.015 - x_4) * (-0.776 + x_5) * x_9) + ((x_7 - x_6) * (x_9 * x_7)))$$

$$F = -\left(\frac{x_6}{\left(((x_3 - x_8) * (x_1 + x_5)) * x_7 + \frac{x_5}{x_9}\right)}\right) - x_1$$

$$Score = A + B + C + D + E + F$$

X0 = number of male audios *X1 = number of female audios*

X2= Voice activity time of male *X3 = voice activity of time of female*

X4 = Standard deviation of Energy levels male *X5= Standard deviation of Energy levels female.*

X6 = Standard deviation of amplitudes male *X7 = Standard deviation of amplitudes female*

X8 = Standard deviation of pitch male *X9 = Standard deviation of pitch female*

Upon extensive evaluation of the bias score generated by the metric, a notable pattern emerged: datasets biased towards male speakers consistently produced positive bias scores, while datasets biased towards female speakers yielded negative scores. This directional sensitivity of the metric was sufficient to indicate the presence and direction of gender bias within a given dataset. However, while the score effectively signaled *whether* a bias existed, it did not inherently quantify the magnitude of that bias in a meaningful or interpretable range.

To address this, the bias score was normalized using min-max scaling, transforming the values to fall within a fixed range of -10 to 10, where +10 would ideally represent complete male bias, -10 complete female bias, and 0 a balanced dataset. This scaling was intended to enhance interpretability and provide a consistent measure of the extent of bias in any given dataset.

However, further testing revealed a limitation in this approach. The scaled scores were accurate and meaningful only up to a bias level of approximately 40% in either direction. Beyond this threshold, the score began to gradually decrease, rather than continuing to increase in alignment with the introduced bias. These findings are visually represented in the graphs provided below, which illustrate the degradation in scoring consistency beyond the 40% bias point.

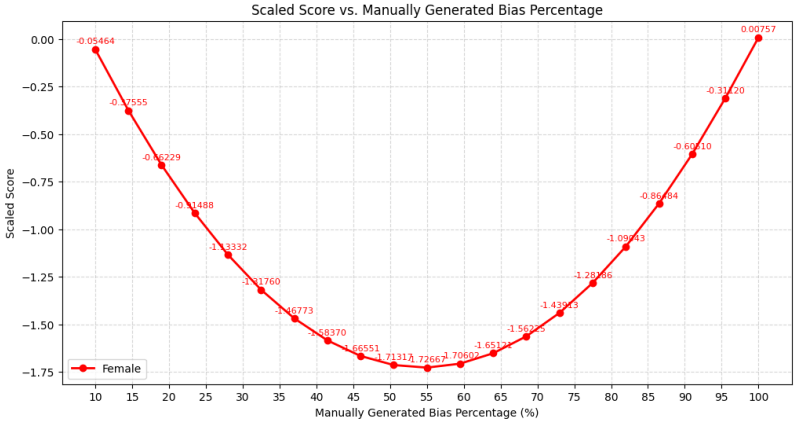

Figure 49 Method- Symbolic Regression : Female Scaled Score Vs Generated bias percentage

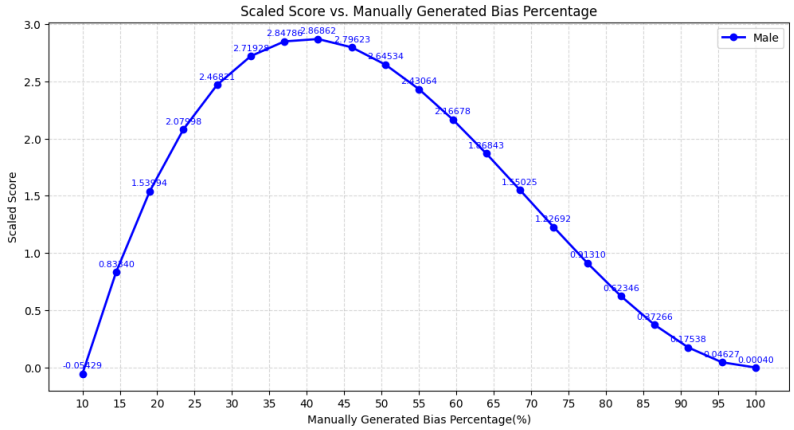

Figure 50 Method- Symbolic Regression : Male Scaled Score Vs Generated bias percentage

As it does not perform the way it is expected to and as the MSE value is significantly higher A new methodology was needed to be studied.

Building the equation: Method - Polynomial Regression with Ridge (L2) regularization.

The augmented dataset was further utilized to train a predictive model using the Elastic Net Regression technique. Elastic Net combines both L1 (Lasso) and L2 (Ridge) regularization, making it well-suited for scenarios where multicollinearity exists and where feature selection is beneficial. To identify the optimal configuration, the model was trained iteratively using a grid search approach across a range of hyperparameters.

Specifically, the regularization strength (alpha) was varied over the set [0.001,0.01,0.1,1.0,10.0][0.001, 0.01, 0.1, 1.0, 10.0][0.001,0.01,0.1,1.0,10.0], while the mixing parameter (l1_ratio) was varied over [0.0,0.1,0.5,0.9,1.0][0.0, 0.1, 0.5, 0.9, 1.0][0.0,0.1,0.5,0.9,1.0], where a value of 0.0 corresponds to pure Ridge regression and 1.0 corresponds to pure Lasso regression.

The best performing configuration was found to be:

- Alpha: 0.01
- L1 Ratio: 0.0 (indicating a purely Ridge-based solution)

This configuration yielded a Mean Squared Error (MSE) of 0.0016 and an R-squared value (R^2) of 0.9998, suggesting a nearly perfect fit to the data with very low prediction error. The final regression equation derived from this model is as follows:

$$\begin{aligned} Bias\ Score = {} & -0.0125 + (0.0001C_{male}) + \left(-0.0001C_{female}\right) + (0.0005V_{male}) \\ & + \left(-0.0005V_{female}\right) + (0.0044E_{male}) + \left(-0.0034E_{female}\right) \\ & + (-0.0004A_{male}) + \left(-0.0004A_{female}\right) + (-0.0334P_{male}) \\ & + \left(-0.0325P_{female}\right) + (0.0002P_{male}^2) + \left(-0.0002P_{male}P_{female}\right) \\ & + (0.0002P_{female}^2 \end{aligned}$$

$C_{male}: Count\ of\ male\ audios$ $C_{female}: Count\ of\ female\ audios$

$V_{male}: Voice\ activity\ male$ $V_{female}: Voice\ activity\ female$

$E_{male}: Standard\ deviation\ of\ energy\ male$

$E_{female}: Standard\ deviation\ of\ energy\ female$

$A_{male}: Standard\ deviation\ of\ Amplitude\ male$

$A_{female}: Standard\ deviation\ of\ Amplitude\ female$

$P_{male}: Standard\ deviation\ of\ Pitch\ male$

$P_{female}: Standard\ deviation\ of\ Pitch\ female$

An extensive testing process was done to analyze how the equation performed on different datasets like with symbolic regression it was found out that when a dataset is biased towards males the score is a positive value and when the dataset is biased towards female the score is a negative value.

Figure 51 Method- Polynomial Regression : Female Scaled Score Vs Generated bias percentage

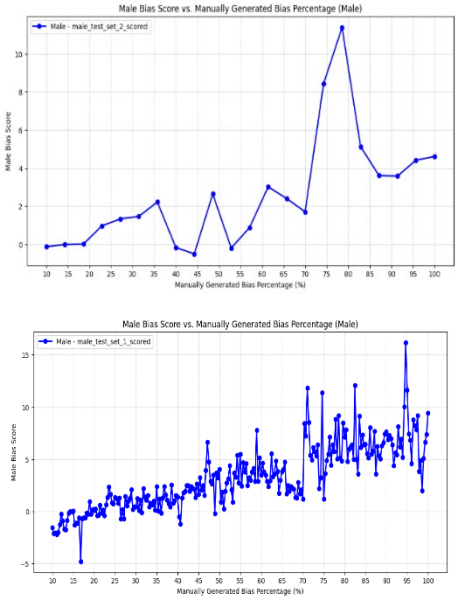

Figure 52 Method- Polynomial Regression : Male Scaled Score Vs Generated bias percentage

The behaviour of the polynomial regression equation was qualitatively similar to symbolic regression with the key difference: the trend of the score. The score calculated using the equation generated through polynomial regression displayed a consistent increase with the increasing levels of bias while the equation generated through symbolic regression displayed a gradual decline beyond a 40% bias threshold despite the increasing levels of bias. Despite the score increasing with the increasing levels of bias the score values could not quantify the degree of bias present in the dataset. To address the limitation min-max scaling was used to scale the score into a more interpretable range.

Based on the understanding of the score it was found, a positive score indicated the dataset to be biased towards male and negative score indicated the dataset to be biased towards females. Based on polarity, the min-max scaling was customized. If the polarity of the calculated score was positive, the maximum bound was derived by assigning the female associated feature values to a minimum and minimum bound was derived by assigning equal values to both the female and male associated features. When the polarity of the calculated score was negative the process was reversed, and corresponding bounds were calculated. The normalization approach performed more consistently across the test datasets.

Through Evaluation of performance of both equations it was found that the equation generated through polynomial regression worked best for the intended metric.

2.3.4. Validation.

To validate the proposed bias metric, we employed the Word Error Rate (WER) — one of the most prominent and widely accepted evaluation metrics in speech recognition and bias detection in audio-based AI systems. WER offers insight into model-based bias, whereas our newly developed Bias Score is designed to capture data-inherent bias, independent of the model's architecture or training process.

For this validation, we used the Whisper-tiny model developed by OpenAI [47]. This model was chosen due to its multilingual support and lightweight nature, making it suitable for consistent testing across diverse datasets and languages. All datasets used for validation were passed through the Whisper-tiny model to obtain WER values for both male and female speakers separately.

A system is considered biased towards a specific gender when its WER is consistently lower for one gender compared to the other. For instance, if the WER is significantly higher for female speakers, the system is said to be male-biased, indicating that the model performs better on male audio data.

On the other hand, the Bias Score is calculated based on statistical audio features (e.g., pitch, energy, speech activity) extracted directly from the dataset, not the model output. A positive Bias Score indicates the dataset is biased towards males, while a negative score implies female bias.

The following table summarizes the WER values per gender, the bias interpretation based on WER, the corresponding Bias Score, and its interpretation. This comparison helps validate whether the Bias Score aligns with the bias suggested by WER:

Datasets	WER male	WER female	Bias defined by WER	Bias score	Definition of the score
LibriSpeech: Dev split	20.05	26.3	System : Male biased	0.650504	Dataset : Male Biased
LibriSpeech: Test-other split	16.0625	19.0588	System : Male biased	1.263191	Dataset : Male Biased
LibriSpeech: Test-clean split	22.45	24.75	System : Male biased	0.811159	Dataset : Male Biased
LibriSpeech Train split	31.5104	32.0797	System : Male biased	0.713163	Dataset : Male Biased
Multi-lingual LibriSpeech : Portuguese	1.148	1.2917	System : Male biased	-4.79006	Dataset : Female Biased
Multi-lingual LibriSpeech : Polish	1.222	1.2009	System : Female biased	6.279849	Dataset : Male Biased
Common voice : Hakha Chin Train split	1.17	1.2114	System : Male biased	1.259123	Dataset : Male Biased
Common voice : Hakha Chin other-split	1.1913	1.1497	System : Female biased	-5.24067	Dataset : Female Biased
Common voice : Hakha Chin validated split	1.1816	1.1599	System : Female biased	-0.23507	Dataset : Female Biased
Common voice : Chuvash test split	1.2571	1.2261	System : Female biased	-3.51048	Dataset : Female Biased
Common voice : Chuvash validated split	1.2433	1.2966	System : Male biased	2.512189	Dataset : Male Biased
Common voice : Sorani validated split	1.2121	1.4281	System : Male biased	8.25787	Dataset : Male Biased
Common voice : Sorani other split	1.1745	1.5312	System : Male biased	8.358887	Dataset : Male Biased

Table 4 WER vs Bias Score

As seen in the table above, in most dataset splits, both the Bias Score and the WER-based interpretation agree on the dominant gender bias. The LibriSpeech splits consistently show a male bias, both in WER and Bias Score. The Sorani and Tartar splits have significantly higher Bias Scores (e.g., >8), which aligns with a system-level male bias indicated by higher female WERs. In a few cases, such as Multilingual LibriSpeech: Polish, there is a contradiction, where WER suggests a female-biased system, but the Bias Score shows male-biased data. These discrepancies may arise from model-specific error patterns or data complexity not captured directly by statistical features.

Such inconsistencies reinforce the idea that while WER is helpful, it is not purely reflective of dataset bias — it also captures model biases, vocabulary familiarity, and architecture sensitivity. In contrast, the Bias Score remains independent of the model and strictly analyzes the data-driven disparities, offering a complementary perspective for gender bias assessment in speech datasets.

2.4. Detecting Gender Bias in Human Activity Video Datasets: A Multi-Component Visual Metric Approach

2.4.1. Overview of the Research Framework

This study introduces a modular, interpretable metric framework for detecting gender bias in human activity video datasets. The proposed methodology addresses the lack of dataset-level fairness evaluation tools for video-based systems by measuring representational disparities across five dimensions, Size, Centering, Screen Time, Embedding, and Motion. Each component captures a distinct visual or motion-based bias signal. The approach encompasses data preparation, deep feature extraction, bias score computation, normalization, and aggregation.

2.4.2. Dataset Description

The dataset used in this research is a curated subset of 500 videos from the Human-Centric Atomic Action (HAA500) dataset [48], which offers high-resolution annotations for over 500 action classes. Each selected video is labeled by activity type and the perceived gender (male/female) of the performing subject. This dataset's diversity in physical activities and framing techniques supports the exploration of spatial and motion-based gender bias.

2.4.3. Justification for Bias Metric Components

The components selected for measuring gender bias in human activity video datasets were chosen based on their empirical grounding in visual bias literature and their feasibility using state-of-the-art computer vision models. Each metric captures a distinct aspect of representational bias, contributing to a holistic and interpretable evaluation framework:

- **Size Bias**: Bounding box size is a well-established proxy for visual prominence. Larger subjects within a frame draw more attention, and consistent prominence of one gender can reinforce biased associations. This form of bias has been linked to uneven representations in tasks like captioning and detection [49].
- **Centering Bias**: Visual salience is often enhanced when subjects are placed near the center of a frame. Gender-based disparities in central positioning have been documented, reflecting subtle framing preferences that influence viewer perception, especially in surveillance footage [18].
- **Screen Time Bias**: Longer visibility of a subject within a video enhances narrative emphasis. Gender imbalance in screen presence can influence both training outcomes in models and viewer interpretation, especially in datasets used for human activity recognition [10].
- **Embedding Bias**: Semantic embedding bias is assessed via similarity of deep video features to gendered activity profiles. These embeddings capture high-level

associations that, if biased, could misrepresent gender roles or behavioral expectations [49].

- **Motion Bias**: Differences in pose dynamics and motion trajectories between genders are crucial in activity classification. Empirical research shows that these stylistic movement patterns can reveal gender tendencies and are effective for bias measurement [50].

These dimensions collectively enable a granular diagnosis of visual and motion-based bias in video datasets and justify the multi-component design of the proposed metric framework.

2.4.4. Visual Feature Extraction Pipeline

Feature extraction was conducted using a three-model pipeline tailored to capture spatial scale, motion trajectories, and semantic embeddings:

Spatial Features via YOLOv8

YOLO (You Only Look Once) is a widely adopted object detection framework designed for real-time performance and high spatial accuracy. In this study, YOLOv8 is employed to detect individuals in video frames and extract the bounding box area (for size bias) and bounding box center coordinates (for centering bias). YOLO predicts object locations in a single forward pass, enabling efficient per-frame analysis of spatial positioning and scale [20].

Motion Features via MediaPipe

Motion bias is quantified by analyzing pose-based movement across video frames. For this, the MediaPipe Pose solution is used, which offers high-fidelity estimation of 33 key points per frame, including body, face, and hand joints. These pose vectors are converted

into normalized motion trajectories, allowing the system to detect stylistic or dynamic differences by gender [52].

Embedding Features via SlowFast

To represent the semantic content of each video, the SlowFast model is used to extract high-level action embeddings. SlowFast operates by processing video at two temporal resolutions: a slow pathway for semantic context and a fast pathway for motion detail. This dual-stream architecture allows for robust feature learning from human-centric actions, which is critical for computing embedding bias using gender-specific centroid similarity [53].

2.4.5. Component-Level Metric Design

To capture different dimensions of gender bias, five individual bias metrics were computed per video. Each metric targets a unique visual or motion-related attribute and is later combined into composite bias scores.

Size Bias

Purpose:

To measure the visual prominence of a subject based on their relative size within the video frame.

Inputs:

- x_1, y_1, x_2, y_2: Bounding box coordinates (from YOLO)
- W, H: Frame width and height

Computation:

1. Bounding box area:

$$A_{bbox} = (x_2 - x_1) * (y_2 - y_1)$$

2. Frame area:

$$A_{frame} = W * H$$

3. Size ratio per frame:

$$Rsize = \frac{A_{bbox}}{A_{frame}}$$

4. Final score (video-level mean with gender sign):

$$Size\ Bias = mean(Rsize) * \begin{cases} +1 : if\ gender\ =\ male \\ -1 : if\ gender\ =\ female \end{cases}$$

Centering Bias

Purpose:

To assess how centrally the subject is framed within the video.

Inputs:

- Bounding box center: $x_c = \frac{(x1+x2)}{2}$, $y_c = \frac{(y1+y2)}{2}$
- Frame center: $f_x = \frac{W}{2}$, $f_y = \frac{H}{2}$

Computation:

1. Euclidean distance to center:

$$d = \sqrt{(x_c + f_x)^2 + (y_c + f_y)^2}$$

2. Normalized centering score:

$$D_{norm} = 1 - \frac{d}{(0.5*\sqrt{W^2+H^2})}$$

3. Final score:

$$Centering\ Bias = mean(D_{norm}) * \begin{cases} +1 : if\ gender = male \\ -1 : if\ gender = female \end{cases}$$

Screen Time Bias

Purpose:

To measure the duration of on-screen visibility for individuals in each video, while incorporating gender directionality. This metric reflects whether a video's subject contributes more to male-leaning or female-leaning representation based on screen exposure.

Inputs:

- f_v: Number of detected frames for a given video
- T_{total}: Total number of detected frames across all videos in the dataset

Computation:

Final score:

$$Screen\ Time\ Bias = \frac{f_v}{T_{total}} * \begin{cases} +1 : if\ gender = male \\ -1 : if\ gender = female \end{cases}$$

Embedding Bias

Purpose:

To evaluate the semantic similarity of each video to male and female action embedding centroids.

Inputs:

- V: Video embedding vector (SlowFast)
- C_{male}, C_{female}: Gender-specific embedding centroids

Computation:

1. Normalize embedding:

$$V' = \frac{V}{||V||}$$

2. Cosine distances:

$$d_m = cosine(V', C_{male}) \quad , \quad d_f = cosine(V', C_{female})$$

3. Final metric:

$$Embedding\ Bias = \ d_f \ - \ d_m$$

Motion Bias

Purpose:

To capture gender differences in movement patterns using normalized pose dynamics.

Inputs:

- P_t : Pose keypoints at time t (from MediaPipe), normalized by frame size
- C_{motion_male}, C_{motion_female} : Gender-specific motion centroids

Computation:

1. Frame-level motion vector:

$M = mean(|P_{t+1} - P_t|)$; across all frames

2. Cosine distances:

$dm = cosine(M, C_{motion\_male})$, $dm = cosine(M, C_{motion\_female})$

3. Final score:

$Motion\ Bias = d_f - d_m$

2.4.6. Normalization of Bias Components

Each of the five individual bias components, Size Bias, Centering Bias, Screen Time Bias, Embedding Bias, and Motion Bias are computed on different scales and units. Without normalization, these raw values would contribute unequally to any aggregated metric, disproportionately emphasizing those with higher variance or broader ranges. For instance, centering bias values may range from −0.9 to +0.9, while screen time bias could span only a narrow band such as −0.01 to +0.01. This imbalance would distort downstream interpretations, particularly in methods like Principal Component Analysis (PCA), where variance magnitude directly affects feature weighting.

Normalization Approach

To ensure fair contribution from each component, all five metrics were standardized using Z-score normalization via *StandardScaler*. This method transforms each feature to have:

- A mean of 0
- A standard deviation of 1

Mathematically, each value x is transformed as:

$$x_{std} = \frac{x - \mu}{\sigma}$$

Where μ is the mean and σ is the standard deviation of the respective bias component. This scaling ensures that all metrics contribute equally in terms of statistical variance, regardless of their original range or distribution.

2.4.7. Metric Aggregation

Once the five individual bias components were normalized using Z-score standardization, three composite metrics were derived per video to capture different perspectives of bias: directionality, intensity, and data-driven weighting. These aggregated scores allow for holistic interpretation and comparison of gender bias across activities.

Directional Score

Purpose:

To summarize the overall gender tilt of a video by averaging its normalized bias components.

Computation:

Let B_1, B_2, B_3, B_4, B_5 represent the standardized values of the five bias components. The directional score is computed as:

$$Directional\ Score = \frac{1}{5}\sum B_i$$

Interpretation:

- Positive scores indicate male-leaning bias
- Negative scores indicate female-leaning bias
- Values close to 0 imply balanced or neutral representation

This metric is simple and interpretable, offering a general view of bias polarity.

Magnitude Score

Purpose: Measures the overall strength of bias, regardless of direction.

Computation:

$$\text{Magnitude Score} = \sum_{i=1}^{5} |B_i|$$

Interpretation:

- Higher scores reflect stronger or more extreme bias
- A value of 0 would indicate complete neutrality across all dimensions

This metric is useful when direction is less important than the extent of distortion in representation.

PCA-Weighted Score

To derive a data-driven bias score that emphasizes the components with the highest variance across the dataset.

Computation:

Principal Component Analysis (PCA) was applied to the standardized metrics, and the first principal component was extracted:

$$PCA\ Score = w_1\mathrm{B1} + w_2\mathrm{B2} + w_3\mathrm{B3} + w_4\mathrm{B4} + w_5\mathrm{B5}$$

Where w_i are the PCA weights derived from component variance.

Rescaling for Interpretability:

Since PCA scores are unbounded, the output values were rescaled to the range [-1,+1] using MinMax normalization:

$$PCA - Weighted\ Score = -1 + 2 * \frac{PCA\ Score - min(x)}{max(x) - min(x)}$$

Interpretation:

- +1: Highest male-leaning bias
- -1: Highest female-leaning bias
- 0: Balanced representation

This score captures the most statistically meaningful pattern of bias present in the data, weighted by natural variability.

2.4.8. Limitations

While the proposed framework provides a systematic and interpretable approach to measuring gender bias in human activity video datasets, several limitations should be acknowledged to contextualize the scope and applicability of the findings.

- **Dataset-Specific Generalizability**

The analysis was conducted using a single video dataset with a fixed structure and predefined activity classes. Although the results are internally consistent, they may not generalize across datasets that vary in cultural context, resolution, or collection methodology. Cross-dataset validation is necessary to confirm the broader applicability of the framework.

- **Binary Gender Constraint**

The current methodology operates under a binary gender classification (male/female), which, while simplifying the computational model, does not account for non-binary, transgender, or other gender-diverse representations. This limitation reflects broader dataset labeling practices and underscores the need for more inclusive annotations in future work.

- **Proxy Indicators of Bias**

Some bias components such as screen time, centering, and bounding box size—are used as proxies for representational prominence. These indicators may not fully capture the sociocultural dimensions of bias or viewer perception. Nevertheless, they provide useful, quantifiable signals that can be aggregated for pattern detection and comparison.

- **Label and Detection Fidelity**

The reliability of bias metrics is directly tied to the accuracy of person detection, pose estimation, and gender labeling. Although robust models (YOLOv8, MediaPipe, SlowFast) were used, errors in key point extraction or gender assignment can introduce noise, particularly in low-light or occluded scenes.

- **Component Independence Assumption**

Each metric was computed independently, assuming orthogonality across spatial, motion, and embedding dimensions. However, real-world biases may involve interdependencies, for instance, motion dynamics may correlate with framing style. While independence simplifies analysis, future extensions could explore feature interactions or multivariate coupling.

3. RESULTS AND DICUSSION.

This section presents a comprehensive analysis of the results obtained from applying the proposed bias detection methodologies across four different data modalities: text, image, audio, and video. Each subsection explores how bias manifests in that particular modality, evaluates the corresponding bias scores generated by the proposed equations, and interprets the results in relation to ground-truth labels or expected distributions. Through these focused evaluations, the section highlights the effectiveness and limitations of the bias detection equations and their ability to capture representational disparities across demographic groups such as gender. The results serve as a foundational step toward developing a unified bias detection metric for multimodal datasets.

3.1. Results And Discussion of Developing a Metric to Detect Gender Bias In Contextual Word Embeddings.

3.1.1. Results

This section presents the experimental results obtained by applying the Context-Aware Bias Metric (CABM) to the curated WinoBias-based dataset using three distinct weighting strategies: PCA, PCA + Random Forest, and SHAP + Random Forest. Results are structured around computed bias scores, statistical validation outputs, visualizations, and category-based evaluations.

CABM Score Distribution

The kernel density estimate (KDE) plots showed how bias scores varied across the three approaches. The SHAP + RF method produced a balanced and stable distribution centered near neutrality, while the PCA-only scores showed a wider spread with weaker bias separation. The PCA + RF distribution closely aligned with SHAP + RF, indicating strong consistency between Random Forest–based approaches (Figure 11) .

- **Observation:** SHAP + RF results demonstrated the most controlled and interpretable distribution of bias scores.

Bias Category Distribution

Occupations were categorized as Male-Biased**,** Female-Biased**,** or Neutral based on CABM scores using a +0.1, -0.1 threshold. Bar chart analysis revealed that SHAP + RF assigned a relatively balanced distribution across all three categories, whereas PCA-only classified a larger portion of occupations as neutral, failing to highlight directional bias. PCA + RF followed a pattern similar to SHAP + RF, with slightly fewer Female-Biased classifications .

Statistical Testing Results

Several statistical validation tests were performed to evaluate the quality and discriminative power of the CABM metric:

- Shapiro-Wilk Test: All three bias score distributions passed the normality test ($p > 0.05$), confirming that scores are statistically valid for further parametric analysis.
- Mann-Whitney U Test: A significant difference was found between Male-Biased and Female-Biased groups under the SHAP + RF method ($U = 198.00$, $p < 0.0001$), demonstrating the ability of CABM to clearly distinguish gendered associations in contextual embeddings (Figure 12).
- Score Stability (Standard Deviation & Variance):SHAP + RF showed the lowest standard deviation (0.378) and variance (0.143), indicating stable and consistent scoring.PCA + RF exhibited the highest variability (std = 0.505, var = 0.255), suggesting greater fluctuation across occupations.
- Discriminative Power (Cohen's d & Median Difference):All methods demonstrated large effect sizes (Cohen's $d > 3.0$), validating CABM's ability to separate male- and female-biased categories.PCA + RF had the highest effect size ($d = 4.14$), while SHAP + RF maintained a strong effect ($d = 3.33$) with more

controlled variability.

- Robustness to Outliers (IQR):SHAP + RF had the lowest interquartile range (IQR = 0.385), suggesting focused central score clustering and less susceptibility to extreme values.PCA + RF had the widest spread (IQR = 0.639), indicating noisier behavior in mid-range predictions.
- Correlation Analysis:

  SHAP + RF vs. PCA + RF: Pearson r = 0.93, Spearman ρ = 0.93

  SHAP + RF vs. PCA Only: Pearson r = 0.70, Spearman ρ = 0.64

  PCA + RF vs. PCA Only: Pearson r = 0.79, Spearman ρ = 0.70

These results highlight high agreement between SHAP + RF and PCA + RF, and weaker consistency with PCA-only, confirming that PCA lacks sensitivity to nonlinear and context-specific bias patterns.

Bias Category Agreement Rates

Agreement rates were calculated by comparing how each method classified occupations:

Comparison	**Agreement (%)**
SHAP + RF vs PCA + RF	64.10%
SHAP + RF vs PCA Only	94.87%
PCA + RF vs PCA Only	69.23%

This strong agreement between SHAP + RF and PCA Only demonstrates that SHAP-based contextual weighting aligns well with traditional methods, while the lower agreement with PCA + RF suggests that Random Forest–based PCA scores may be less sensitive to directional bias variations.

3.1.2. Research Findings

The results from applying the Context-Aware Bias Metric (CABM) across the dataset revealed several important insights into how gender bias manifests in contextual word embeddings, as well as how different weighting strategies influence detection quality.

Contextual Bias is the Dominant Signal

The SHAP-based feature importance analysis demonstrated that sentence-level contextual bias was consistently the strongest contributor to the final CABM scores. This confirms the central hypothesis of the study: that contextual information such as how a word like *"nurse"* is embedded in sentences with *"he"* versus *"she"* captures a deeper layer of bias than purely semantic (cosine similarity) or frequency-based (PMI) methods.

This finding supports the growing consensus in NLP fairness research that bias is not static, but often arises from context and interaction between words.

Bias Category Distribution

Using a fixed threshold (+0.1 / –0.1), occupations were categorized as Male-Biased, Female-Biased, or Neutral. SHAP + RF showed a confident distribution with a clear spread across all three categories. PCA + RF assigned more occupations as Neutral, reflecting a conservative classification pattern. PCA Only produced more Female-Biased classifications and showed a distribution similar to SHAP + RF.

Random Forest–Based Weighting Provides Stability

Statistical tests confirmed the robustness of the SHAP + RF and PCA + RF weighting methods. All three CABM variants passed the Shapiro-Wilk Test for normality ($p > 0.05$), confirming the distributions were valid for parametric analysis. Additionally, the Mann-Whitney U Test showed a statistically significant difference between Male-Biased and

Female-Biased occupation scores under the SHAP + RF method ($U = 198.00$, $p < 0.0001$), highlighting its strong discriminative ability.

Metric Captures Subtle and Complex Bias Patterns

Unlike existing metrics that focus solely on word-pair analogies or direct associations, CABM was able to highlight **sentence-level nuances**. For instance, occupations like *"teacher"* and *"doctor"* showed small cosine similarities, but significantly higher contextual bias when analyzed across gendered sentence templates.

This finding indicates that CABM can detect **implicit and indirect biases**, which are often overlooked in traditional static embedding analyses.

Unmasking Validation

validation using unmasking predictions from a fine-tuned BERT model further confirmed the strength of SHAP + RF. By comparing predicted pronouns to CABM-generated bias categories, SHAP + RF showed a 69.23% agreement rate substantially higher than PCA + RF (51.28%) and PCA (56.41%). This confirms that SHAP-weighted CABM scores are not only statistically and visually valid, but also closely mirror actual model predictions in biased contexts.

Comparability with Existing Metrics

To assess the robustness of CABM in comparison to established bias detection methods, a controlled experiment was conducted using the Word Embedding Association Test (WEAT) as a baseline. Following the methodology of Schröder et al., both CABM and WEAT were evaluated across three BERT models trained under identical gender distributions but with differing sentence structures (pretrained, biased, and balanced fine-tuned versions).

The standard deviation of bias scores was computed per occupation across the three models to quantify how consistently each metric measured bias.

Metric	Avg. Std. Deviation	Interpretation
WEAT	0.0111	High numerical stability, limited context sensitivity
CABM	0.0385	Slightly higher variance, but more context-aware and expressive

This multi-angle validation confirms that the SHAP + RF-based CABM method is the most expressive, interpretable, and context-aware approach. It demonstrated statistically validated score separation, the strongest agreement with unmasking predictions, balanced bias categorization, and strong alignment with other stable scoring methods. Furthermore, its integration of SHAP-based feature attribution offers direct insight into score composition, supporting its use as the final recommended configuration for contextual gender bias detection in word embeddings.

3.1.3. Discussion

The findings from this study offer compelling evidence for the presence of contextual gender bias in transformer-based word embeddings and demonstrate the effectiveness of the Context-Aware Bias Metric (CABM) as a tool for measuring it. This section discusses the broader implications of the results, addresses the limitations of the current implementation, and reflects on the potential impact of the CABM framework in both academic and real-world applications.

CABM as a Context-Sensitive Metric

Unlike traditional metrics such as WEAT or direct cosine comparisons that rely on static word-level associations, CABM incorporates multiple dimensions semantic similarity,

statistical co-occurrence, and sentence-level embedding shifts. This multi-faceted approach allows CABM to capture deeper, more nuanced forms of bias, especially those that emerge only when a word is placed in different gendered contexts.

The dominance of contextual bias in SHAP analyses confirms that bias in language models is often subtle and embedded in the way models interpret relationships between words, rather than in the words themselves. This insight supports a paradigm shift in bias detection research from static vector comparisons to dynamic, context-aware evaluation frameworks like CABM.

Model Sensitivity and Generalizability

The CABM framework proved to be sensitive to known stereotype patterns. Occupations traditionally associated with men (e.g., *engineer*, *doctor*) showed strong male bias, while roles like *nurse* or *teacher* were more likely to align with female contexts. These results closely mirror societal patterns, suggesting that pretrained language models may unintentionally reinforce harmful gender associations if not evaluated properly.

The application of CABM across three distinct weighting strategies further demonstrated the flexibility of the metric. While SHAP + Random Forest yielded the most interpretable and consistent results, the inclusion of PCA and PCA + RF variants allowed for cross-method comparison and validation. This extensibility makes CABM adaptable for integration into a variety of model types and fairness auditing systems.

3.1.4. Limitations and Future Directions

Despite its strengths, the current implementation of CABM has several limitations:

- The synthetic bias injection test using MLM fine-tuning was not conclusive due to limited dataset size and training instability. While conceptually sound, this experiment requires further refinement with larger, better-structured data to fully

test CABM's sensitivity to known biases.

- CABM is currently designed for binary gender evaluation. Extending the framework to non-binary or intersectional identities will be a critical future step.
- This study focuses on BERT embeddings; while CABM is model-agnostic, further validation across other architectures (e.g., RoBERTa, GPT, XLNet) is needed.

Broader Impact

CABM introduces a context-aware, explainable, and extensible framework for evaluating gender bias in modern NLP systems. As large language models become deeply integrated into public and private sector applications, metrics like CABM can play a vital role in ensuring responsible, transparent, and fair AI development. Its potential for commercialization and open-source deployment further increases its value as a practical, real-world solution.

3.2. Results And Discussion of Developing Metrics for Detecting Gender Bias in Image Datasets Using Contextual Factors: Objects, Scenes, And Spatial Relationships

3.2.1. Results

Experimental Setup Recap

This study was designed to validate a novel metric Unified Bias Metric (UBM) for detecting contextual gender bias in image datasets by analyzing the relationships between detected persons, objects, and scenes. The experimental setup incorporated nine computational approaches applied across eight uniquely structured datasets to ensure a robust evaluation of contextual bias patterns. The entire pipeline was modular, scalable, and reproducible, implemented in Google Colab using Python 3.10+.

Dataset Overview

Dataset	Description
Original Dataset	Full set of 852 annotated samples containing person-object-scene triplets.
Male Biased Dataset	Skewed toward male-labeled images (~70–90%) to test sensitivity to male-prevalent environments.
Female Biased Dataset	Contains a higher proportion of female-labeled images (~70–90%).
Balanced Bias Dataset	Equal male and female distribution across all object classes. Used for baseline fairness evaluation.

Neutral Dataset	Composed of objects equally occurring across genders with no prior known bias.
Male Biased (All Genders)	Maintains all object classes but increases male-label frequency, allowing mixed but skewed bias testing.
Female Biased (All Genders)	Same as above but skewed toward female label distribution.
Top Extreme Bias Dataset	Contains only the most gender-polarized object types (e.g., *handbag*, *baseball bat*), stress-testing bias metrics.

All datasets were stored as .csv files with the following columns:

Gender, Object_Class, Relative_Size, 3D_Distance, Normalized_Depth, Scene_Similarity_Bias

Approach Overview

A total of **nine approaches** were developed to calculate **UBM (Unified Bias Metric)** using different combinations of weighting strategies and statistical models:

Approach	**Method**	**Model Type**
1	PCA / Random Forest Weighting	Regression
2	SHAP Interpretability	Random Forest Classifier
3	SHAP + PCA (Averaged)	RF + PCA Combination
4	SHAP Only	XGBoost Classifier
5	SHAP + SMOTE	XGBoost Classifier (balanced data)
6	SHAP + PCA + Grid Search	XGBoost + SMOTE + Search
7	SHAP + Optuna	XGBoost + Hyperparameter Tuning
8	Ridge Regression	Linear Model
9	SHAP + PCA → Ridge	Final Hybrid Approach

Each approach outputs:

- **Bias Score per object class**
- **Bias Category:** Male-Biased / Female-Biased / Neutral

- **Feature weights** (SHAP/PCA/Ridge)
- **Visual plots** (KDE, bar, SHAP importance)

Execution Setup

- Platform: Google Colab
- Environment: Python 3.10+, GPU/TPU-enabled
- Core Libraries: shap, xgboost, optuna, pandas, matplotlib, seaborn, scikit-learn, imblearn
- Pipeline Script: pipeline_for_9_datasets.py
- Validation Script: validate_the_approaches.py

Each dataset–approach pair was processed via a modular pipeline, storing results in:

/content/drive/MyDrive/ContextualBias/Analysis/results/<dataset_name>/<approach_name>/

This structure ensured full reproducibility, parallel evaluation, and ease of visual analysis across 72 executions (9 approaches × 8 datasets).

3.2.2. Bias Score Distributions Across Approaches

To evaluate the consistency, sensitivity, and robustness of the Unified Bias Metric (UBM), this section presents a comparative distributional analysis of bias scores produced by each of the nine computational approaches across six representative datasets. Each approach outputs a normalized bias score ranging from −1 (female-biased) to +1 (male-biased), with 0 representing neutrality. The underlying formula remains constant, but the scoring weights (derived via PCA, SHAP, Ridge, or XGBoost) introduce variance in how contextual features influence the final score.

Below figure shows boxplots for all nine approaches applied to the **Original Dataset**, illustrating the spread, outliers, and central tendencies of the bias scores:

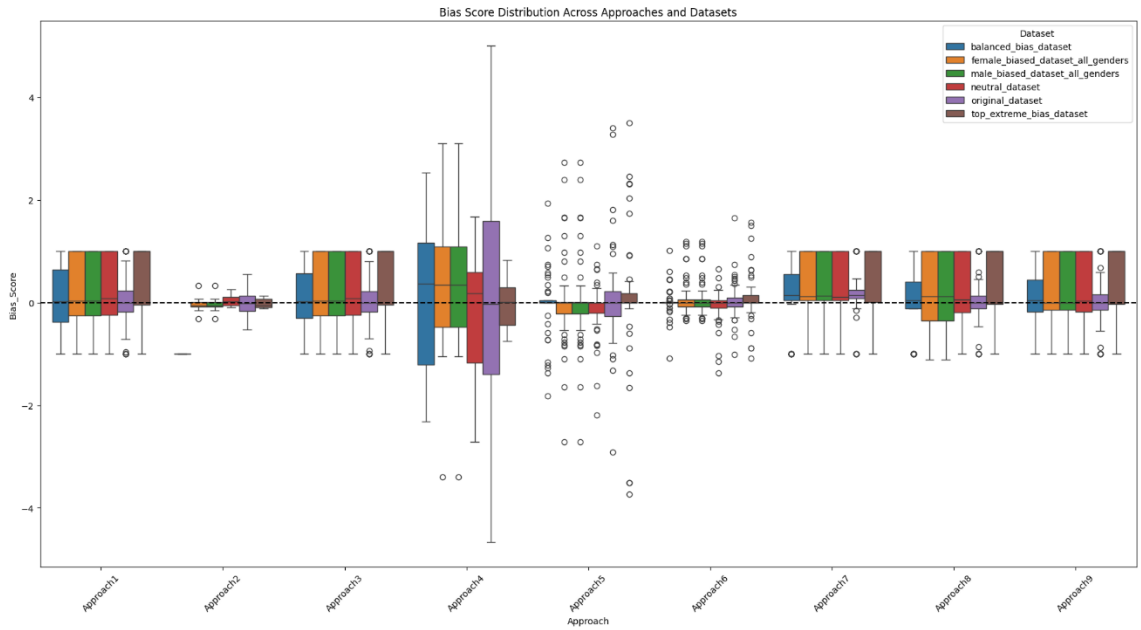

Figure 53Boxplot of Bias Score Distributions Across Approaches (Original Dataset)

- Approaches **1, 2, 3, 6, 7**, and **9** show compact interquartile ranges with balanced medians near zero, suggesting consistency and low bias volatility.
- **Approach 4**, which relies solely on XGBoost and SHAP without PCA or tuning, exhibits high variance and frequent outliers, indicating sensitivity to data noise or feature imbalance.
- **Approaches 5 and 8** demonstrate moderate spread, showing that SMOTE balancing (Approach 5) and Ridge regularization (Approach 8) offer controlled yet flexible scoring.

Below figure displays Kernel Density Estimation (KDE) overlays of the bias score distributions for each approach across six datasets:

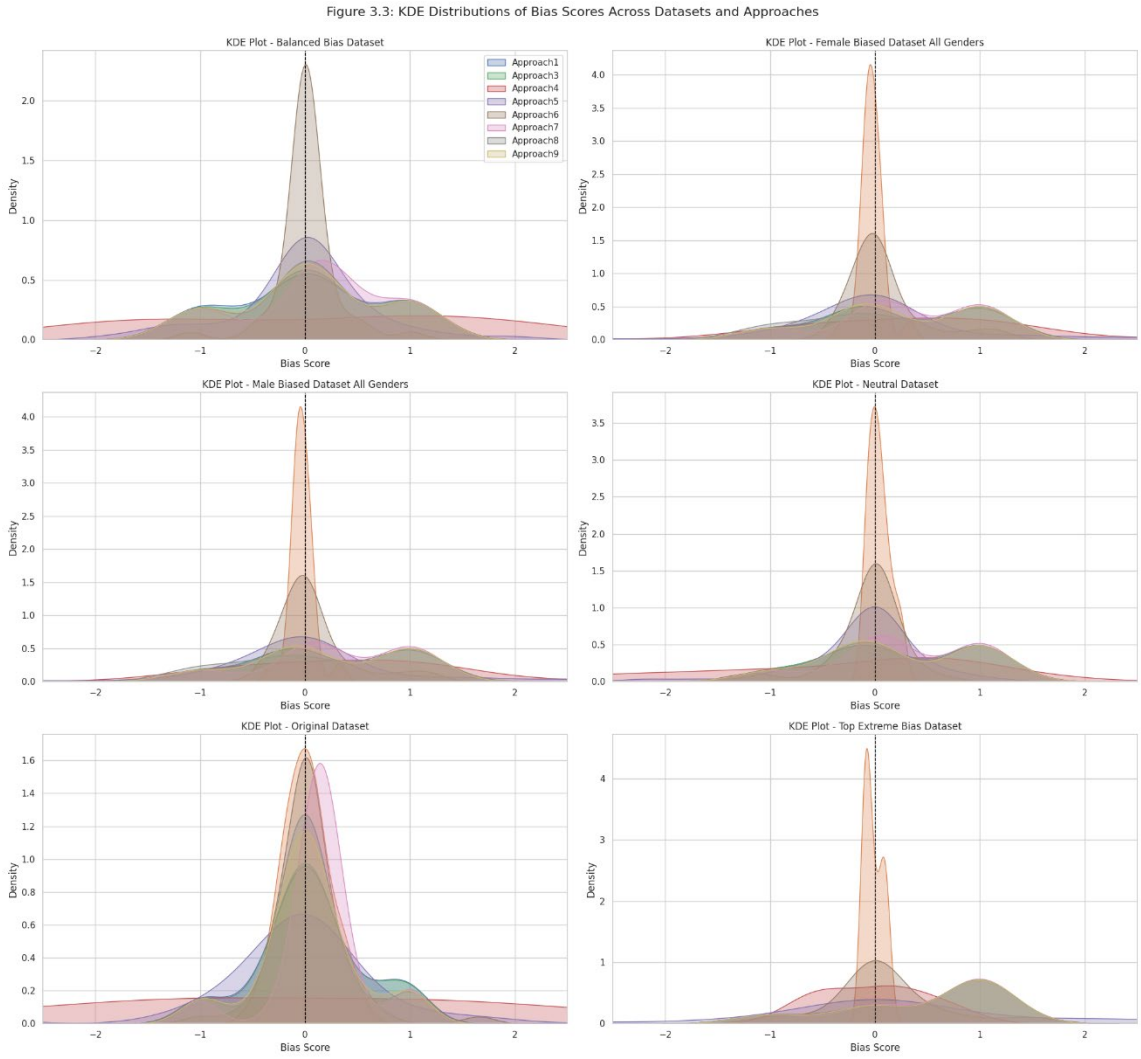

Figure 54 KDE Distributions of Bias Scores Across Datasets and Approaches

- The **Balanced Bias Dataset** exhibits tight central peaks across most methods, indicating fair behavior under controlled distributions.
- **Female- and Male-Biased Datasets (All Genders)** show asymmetric shifts in the KDE curves, validating each method's sensitivity to dataset skew.
- **Neutral Dataset** yields tall, narrow distributions centered around zero, as expected when bias signals are minimized.
- **Top Extreme Bias Dataset** displays broad distributions with heavier tails, consistent with datasets designed to amplify object-gender associations.

- The **Original Dataset** produces mixed-width KDEs, reflecting real-world bias complexity.

Key Insights

- Approaches 6 and 9 (Grid Search + SHAP + PCA hybrids) consistently exhibit narrow, symmetric distributions across datasets, making them strong candidates for stable bias measurement.
- Approach 3, which averages SHAP and PCA weights, also demonstrates excellent generalization.
- The combination of boxplots and KDEs gives both statistical and intuitive insight into how different algorithms interpret gender bias.

This comparative distribution analysis confirms that optimized hybrid approaches like **Approach 6 and Approach 9** provide the most reliable and balanced bias scores. It also highlights the need for caution with under-regularized models (e.g., Approach 4), which may exaggerate bias signals due to high feature sensitivity or lack of tuning.

3.2.3. Dataset-Wise Behavior Across Bias Types

To evaluate the adaptability and robustness of each bias detection approach, we conducted a dataset-wise comparative analysis across six key datasets → Balanced, Top Extreme Bias, Neutral, Original, Female-Biased (All Genders), and Male-Biased (All Genders). Each dataset was selected to simulate distinct contextual dynamics involving object-gender-scene interplay, providing a landscape to stress-test method behavior under varied bias conditions.

Below figure presents a grouped boxplot illustrating the distribution of bias scores across all nine approaches.

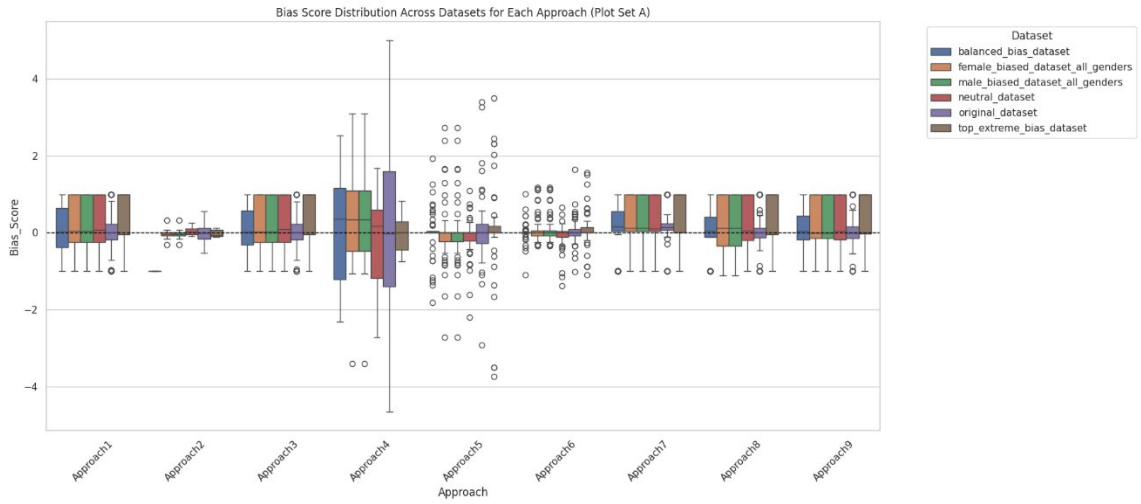

Figure 55 Boxplot Distribution Across Approaches

The analysis reveals that:

- **Approach 2** consistently centers around zero with minimal variance, indicating conservative scoring but poor responsiveness to subtle contextual shifts.
- **Approach 4**, based on SHAP-only XGBoost, demonstrates large variance and significant outlier activity—especially in the original and extreme datasets—suggesting volatility in unregularized model configurations.
- **Approach 6**, leveraging SHAP + PCA with Grid Search, yields the most stable bias score distributions with controlled spread, making it a strong candidate for generalizable deployment.

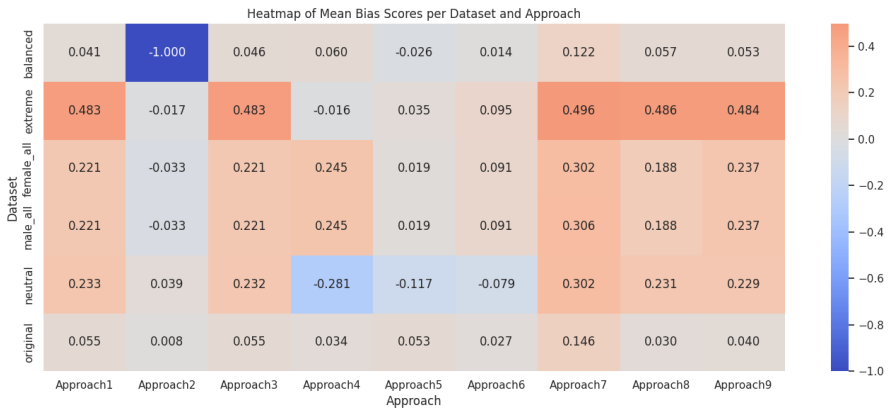

Figure 56: Mean Bias Score Analysis

As shown in the heatmap of average bias scores, several patterns emerge:

- **Approaches 1, 3, 7, and 9** produce consistently positive means across datasets, affirming their sensitivity to male-coded object-scene environments.
- **Approach 4** stands out for its negative mean score in the neutral dataset (−0.281), suggesting either overcorrection or model instability in balanced settings.
- **Approach 7** exhibits strong responsiveness, with mean scores scaling appropriately between balanced (~0.12) and highly skewed datasets (~0.49).

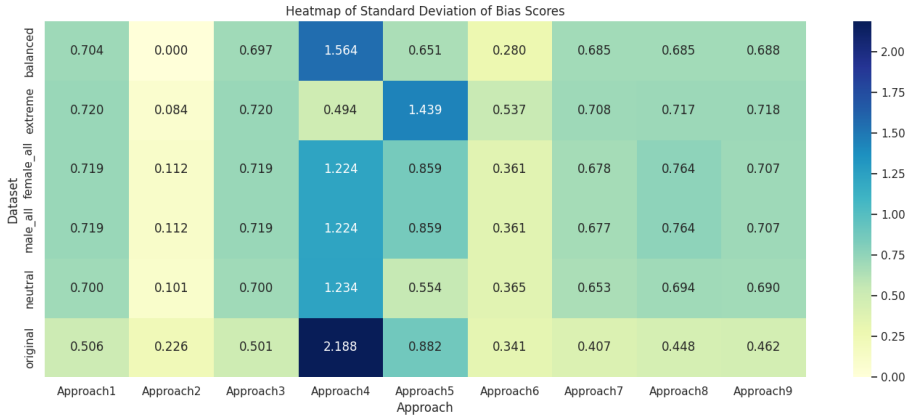

Figure 57:Variance Analysis

Standard deviation values in Figure 3.5 further highlight the reliability of each approach:

- **Approach 4** again emerges as the least stable, showing the highest variance in most datasets, peaking at σ = 2.188 in the original dataset.
- Conversely, **Approaches 6 and 7** maintain low-to-moderate variance across conditions, reinforcing their robustness.
- While **Approach 2** shows minimal variance, it offers limited practical value due to its lack of discrimination between bias types.

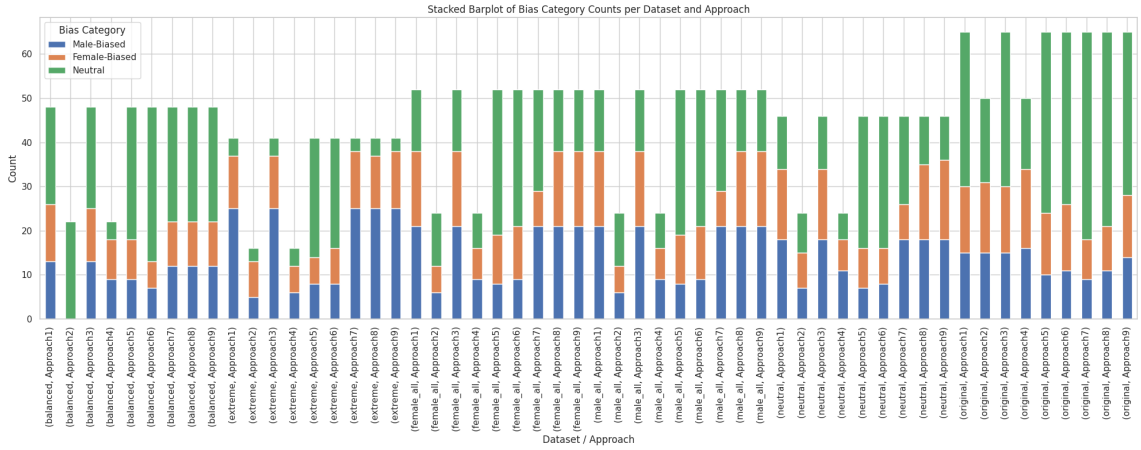

Figure 58: Bias Categorization Trends

The stacked bar plot (Figure 3.6) visualizes how each method classifies objects into Male-Biased, Female-Biased, or Neutral:

- As expected, balanced and neutral datasets yield more neutral classifications across methods.
- In contrast, **Approaches 7–9** successfully identify more male-biased cases in the **Top Extreme Bias Dataset**, aligning with known object-gender stereotypes.
- **Approach 2**, once again, lacks categorical differentiation producing predominantly neutral outcomes even in strongly skewed data, limiting its diagnostic value.

This cross-dataset analysis highlights significant differences in how each computational approach detects contextual gender bias. Combined techniques such as **Approach 6 (SHAP + PCA + Grid Search)**, **Approach 7 (SHAP + Optuna)** and **Approach 9 (SHAP + PCA+ Ridge)** consistently balance sensitivity and stability, outperforming both under-

regularized (e.g., Approach 4) and under-sensitive (e.g., Approach 2) configurations. These observations will guide the final selection of a robust, interpretable UBM strategy in the concluding chapter.

3.2.4. Feature Weighting and Interpretability

To better understand the contribution of individual contextual features in detecting gender bias, this section compares the feature weighting mechanisms employed in three distinct approaches: **Approach 2 (SHAP)**, **Approach 3 (PCA)**, and **Approach 9 (Combined Weights)**. These methods utilize different interpretability frameworks to assign relative importance to three core features:

- **3D Distance** – The spatial separation between the person and object,
- **Normalized Depth** – The object's positioning in 3D space, and
- **Relative Size** – The visual prominence of the object with respect to the person.

Figure 14 illustrates the importance weights derived from each method. The SHAP-based model (Approach 2) distributes relatively low but consistent weights across all features 3D Distance: 0.067, Normalized Depth: 0.046, and Relative Size: 0.039 suggesting a balanced yet conservative attribution of influence.

In contrast, PCA (Approach 3) assigns a dominant weight to 3D Distance (0.861), implying that variance in spatial relationships is the most explanatory feature in the dataset. However, this variance-based emphasis may lead to overfitting, as it downplays the contribution of other contextual cues.

The Combined approach (Approach 9) integrates both SHAP and PCA perspectives likely through normalized averaging resulting in a more evenly distributed set of weights: 3D Distance: 0.409, Normalized Depth: 0.293, and Relative Size: 0.298. This mitigates the overemphasis seen in PCA while enhancing the interpretability provided by SHAP. Notably, the increased weights for Normalized Depth and Relative Size in this hybrid strategy suggest that these features play a more significant role in bias detection than PCA alone would indicate.

This comparison highlights how interpretability methods shape feature prioritization. While SHAP offers model explainability and PCA reveals dominant variance patterns, their fusion in Approach 9 yields a more robust, interpretable, and generalizable weighting scheme for contextual bias detection.

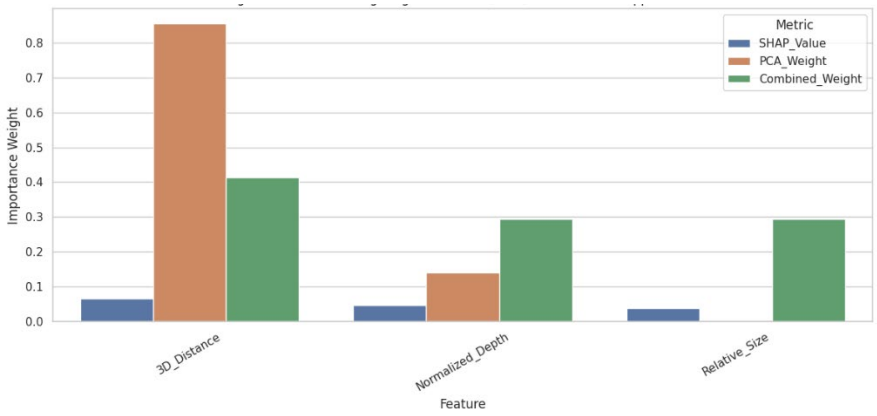

Figure 59:Feature Weighting Across SHAP (Approach 2), PCA (Approach 3), and Combined Weights (Approach 9), showing relative importance assigned to 3D Distance, Normalized Depth, and Relative Size in contextual bias computation.

3.2.5. Statistical Significance Validation

To assess the robustness of the gender-specific bias scores generated by each approach, the Mann–Whitney U test was applied across all 72 combinations of approaches and datasets (9 approaches × 8 datasets). This non-parametric test is well-suited for evaluating differences between two independent groups here, male- and female-labeled objects without assuming a normal distribution of the data.

Using a standard significance threshold of $p < 0.05$, the test evaluated whether the observed bias scores differed meaningfully between gender groups. The results confirmed that all 72 combinations demonstrated statistically significant differences in bias scores, indicating that the detected biases are not random fluctuations but reflect consistent, systematic disparities in person–object–scene relationships.

Importantly, this pattern of significance was maintained across all datasets, including both synthetically biased datasets and the original dataset. Among the approaches, Approach 9 (SHAP + PCA → Ridge) demonstrated particularly strong and consistent statistical validation. On the original dataset, it achieved a p-value of 0.000007, underscoring its sensitivity and reliability in capturing genuine bias patterns in real-world distributions.

These findings validate the Unified Bias Metric (UBM) framework's capacity to detect meaningful gender-based contextual bias. The consistency of statistical significance across diverse data conditions reinforces Approach 9 as one of the most generalizable and interpretable solutions, suitable for deployment in fairness-critical visual AI pipelines.

3.2.6. Research Findings

Validation-Based Comparison of Approaches

To assess the reliability, robustness, and interpretability of each proposed bias detection approach, multiple validation mechanisms were employed. These include alignment with gender misclassification trends, bias category coverage, stability of bias score distributions, cross-approach correlations, divergence analysis, and qualitative alignment with human perception.

Agreement with Misclassification Trends

Bias scores were evaluated for their alignment with full-image gender misclassification data. As shown in **Figure 3.8**, Approaches 9, 8, and 3 achieved the highest agreement with the misclassification bias direction, indicating strong contextual alignment with real-world classification errors. This alignment provides critical validation that the bias scores are not arbitrary but instead reflect actual behavior in downstream tasks.

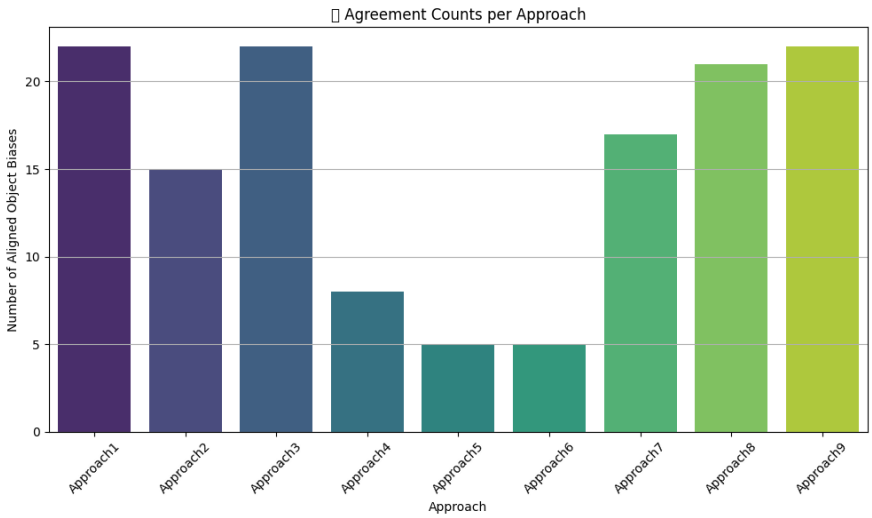

Figure 60:Agreement between predicted object-level bias and full-image gender misclassification trends.

Bias Category Coverage

Coverage in this study refers to the percentage of unique object categories in each dataset for which a method was able to produce a final bias score whether male-biased, female-biased, or neutral. High coverage ensures that no object class was skipped, allowing for complete fairness assessment.

As shown in **Figure 3.9**, **Approaches 1, 3, and 5–9** achieved 100% coverage across all datasets, indicating their robustness in consistently generating scores for every detected object type. In contrast, Approaches 2 and 4 failed to compute scores for a subset of object

classes resulting in lower coverage rates (~39–46%). This omission limits their usefulness in dataset-wide bias analysis.

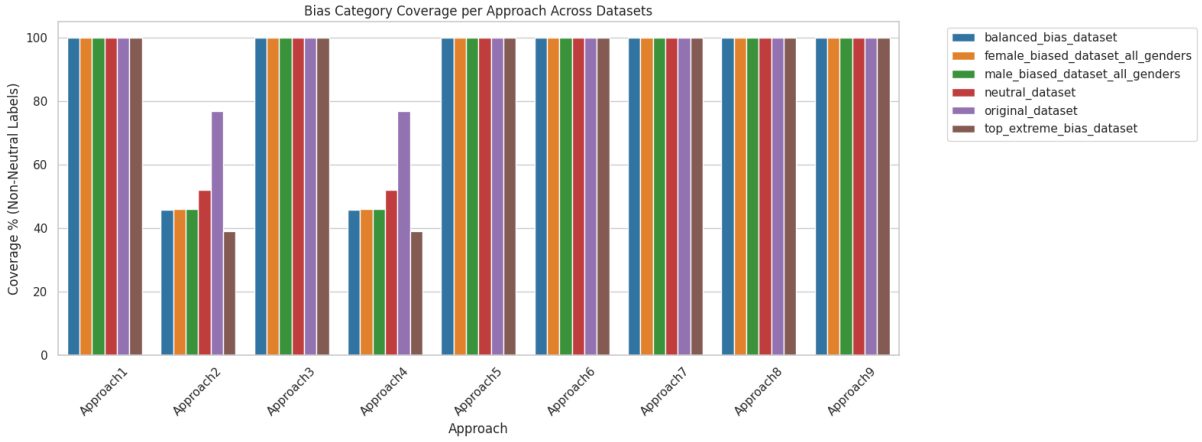

Figure 61: Proportion of unique objects per dataset for which each approach computed a final bias score. Higher values indicate full object-wise coverage, regardless of the final bias category (male, female, or neutral).

Bias Category Distribution

To evaluate the stability and balance of predictions, the distribution of male-biased, female-biased, and neutral labels was analyzed. **Figure 3.10** presents a stacked bar chart showing bias label counts. Approaches 5 and 6 showed an overuse of neutral labels, potentially underestimating gender bias. Approaches 3 and 9 produced well-balanced distributions, indicating consistent and interpretable bias scoring.

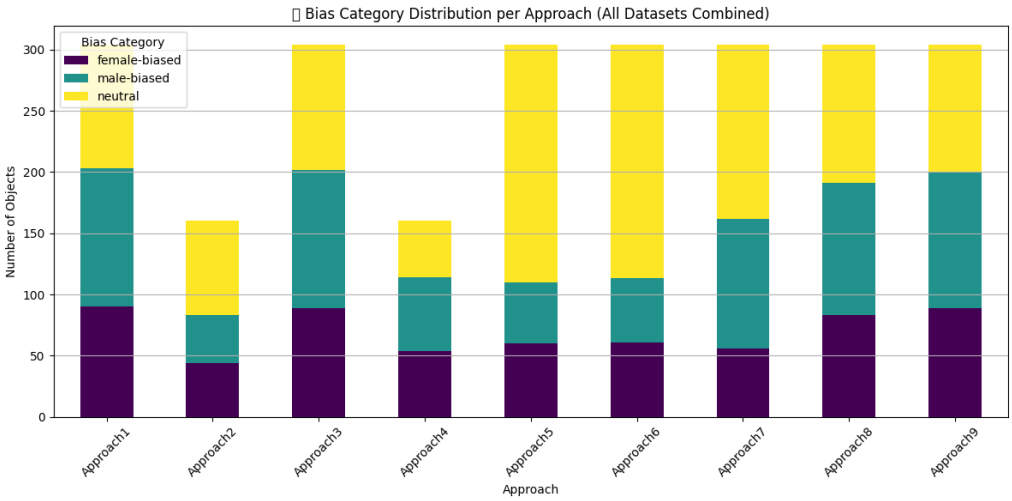

Figure 62: Distribution of bias categories (male-biased, female-biased, neutral) across all datasets.

Cross-Approach Correlation Heatmaps

Correlation analysis was conducted to assess how similar each approach's bias scores were. **Figure 3.11a** shows the **Pearson correlation**, and **Figure 3.11b** shows the **Spearman rank correlation**. Strong correlations were observed between Approaches 1, 3, 8, and 9, suggesting consistent bias patterns. Approach 4 remained an outlier with low correlations, indicating instability and divergence.

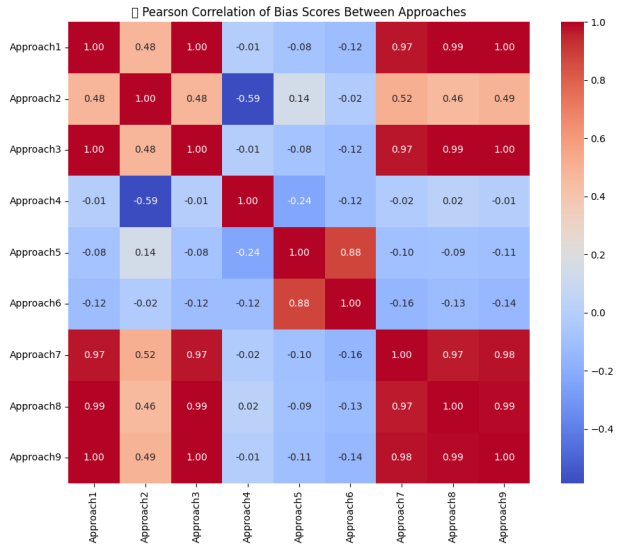

Figure 63 : Pearson correlation heatmap of bias scores across all approaches.

Figure 64 Spearman rank correlation heatmap of bias scores across all approaches.

Cross-Approach Clustering (Dendrogram)

Using hierarchical clustering on the correlation matrix, **Figure 3.12** shows how different approaches group together. Approaches 1, 3, 8, and 9 formed a tight cluster, further supporting their mutual agreement and score consistency. In contrast, Approach 4 formed a separate branch due to its divergent behavior.

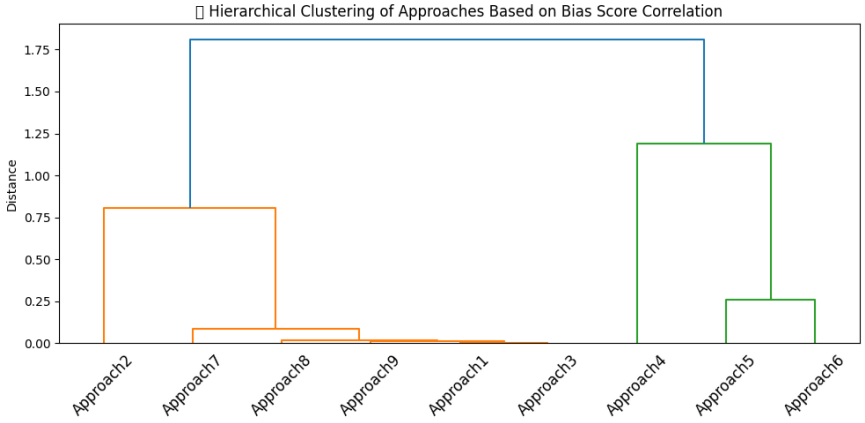

Figure 65: Dendrogram showing hierarchical clustering of bias detection approaches based on correlation similarity.

Top Diverging Object Bias Scores

To further explore inconsistency, the top 10 objects with the most variance in bias scores were identified and visualized in **Figure 3.13**. Approach 4 again exhibited high divergence, assigning extreme scores while others remained near neutral. This confirms instability and over-sensitivity in its scoring logic.

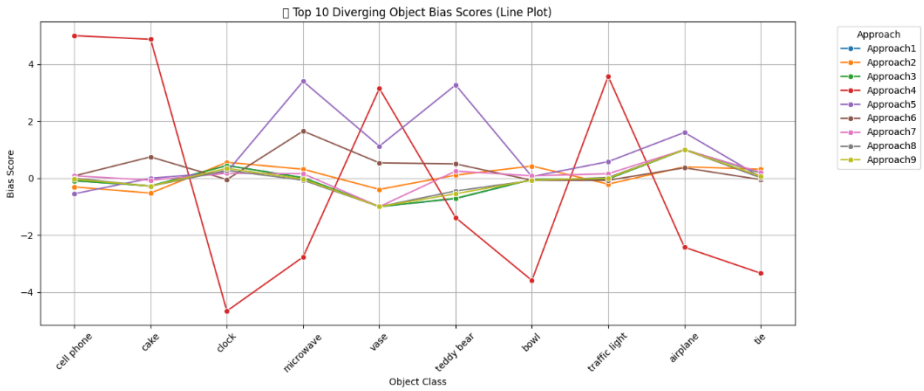

Figure 66: Line plot of bias scores for the top 10 most diverging objects across approaches.

Human Bias Perception Survey

As part of human-centered validation, a survey was conducted where participants were asked to label objects as “typically male,” “typically female,” or “neutral.” **Figure 3.14** shows the aggregated results, sorted by perceived bias. The top human-labeled biases correlated most strongly with Approaches 9 and 3, supporting their real-world validity.

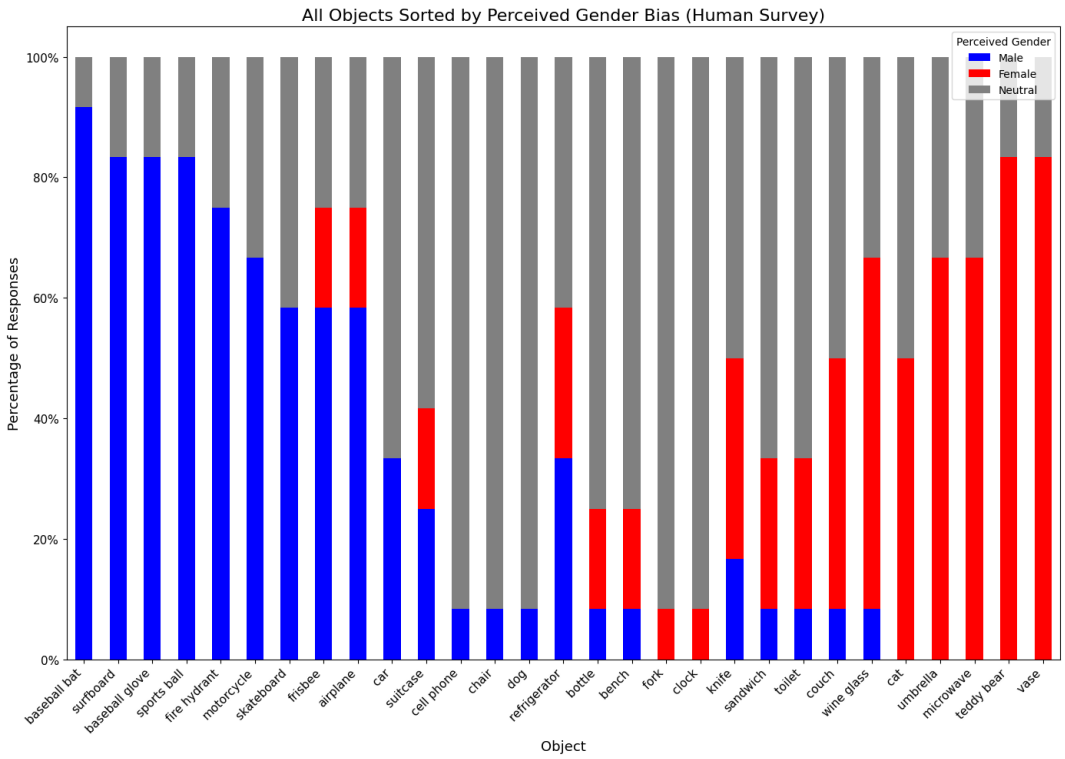

Figure 67: Human perception of gender bias across object classes.

This multifaceted validation demonstrates that **Approach 9 (SHAP + PCA → Ridge)** is the most stable, interpretable, and generalizable across all evaluation metrics. Its high agreement with misclassification trends, complete coverage, balanced label distribution, strong correlation with other stable methods, and high human validation support its use as the final recommended method.

3.2.7. Discussion

Cross-Approach Stability

A core criterion for evaluating bias detection methods is the stability and consistency of their outputs across datasets and evaluation conditions. In this study, a subset of approaches demonstrated strong alignment in both bias score distributions and label assignment patterns, signaling methodological robustness and enhanced reliability for deployment in real-world contexts.

As illustrated in the Pearson (Figure 3.11a) and Spearman (Figure 3.11b) correlation heatmaps, Approaches 1, 3, 8, and 9 exhibited high pairwise correlations, with coefficients frequently exceeding 0.95. This level of correlation indicates that these models consistently assign similar object-level bias scores across datasets, suggesting a shared understanding of contextual bias despite differing computational mechanisms.

This relationship is further substantiated by the hierarchical clustering dendrogram (Figure 3.12), which places these approaches in a tightly linked cluster, reflecting not only numeric similarity but shared scoring behavior. The minimal inter-branch distance reinforces the consistency of their bias detection logic.

These convergences highlight a core group of stable and trustworthy methods particularly Approach 9 that generate reproducible, correlated, and interpretable outputs. Their mutual agreement serves as a form of internal validation, reducing the likelihood of outlier-driven distortions and enhancing confidence in their use for fair and reliable bias auditing.

Limitations of Detection

While the Unified Bias Metric (UBM) framework has demonstrated strong performance across diverse datasets and evaluation strategies, specific computational approaches exhibited notable limitations that merit critical reflection. Identifying these shortcomings is essential for responsible deployment and continuous improvement of bias auditing tools.

Over-Neutral Behavior in Approaches 5 and 6

Approaches 5 (SHAP + SMOTE) and 6 (SHAP + PCA + SMOTE) displayed a marked tendency to over-label objects as “Neutral,” even in datasets with pronounced gender skew. As evident in the bias category distribution plot (Figure 3.10), these methods suppressed both male- and female-biased predictions. While such conservatism might reduce false positives, it risks under-detecting legitimate bias signals particularly in edge-case scenarios like the top_extreme_bias_dataset or the neutral_dataset. This trade-off highlights the need for recalibrated thresholds or adaptive sensitivity tuning.

Instability in Approach 4

Approach 4 (SHAP Only with XGBoost) emerged as an outlier across multiple validation axes. It showed weak correlations with all other methods (Figures 3.11a–b), assigned erratic bias scores to key objects (Figure 3.13), and failed to cluster with stable approaches (Figure 3.12). This instability is likely rooted in overfitting or insufficient regularization in the XGBoost architecture. Although the SHAP values enhance interpretability, the volatility of output reduces its reliability for consistent bias detection.

Context-Agnostic Labeling in Approach 2

Approach 2, based on SHAP explanations from a Random Forest classifier, frequently did not assign any bias score to many object categories across datasets such as balanced_bias_dataset and female_biased_dataset. This behavior resulted in low object-wise score coverage rates (Figure 3.9), where a significant portion of detected object categories were skipped entirely.

Unlike other methods, this conservative scoring likely stems from overly cautious feature attribution or the Random Forest's simplistic decision boundaries. Consequently, subtle contextual biases may go undetected, limiting the model’s utility in thorough fairness assessments.

Binary Gender Labeling as a Structural Constraint

Finally, a shared limitation across all approaches is the reliance on binary gender categorization (Male/Female). While methodologically convenient and aligned with existing dataset annotations, this binary framing excludes non-binary, fluid, or intersectional gender identities. Such simplification constrains the metric’s inclusivity and limits its relevance to broader societal applications. Future iterations of UBM should consider integrating more nuanced gender representations to foster equity in gender-sensitive AI systems.

Interpretation of Bias Trends

A central objective of this research was not only to detect gender bias in image datasets but also to understand how such biases manifest through the interplay of object categories, scene contexts, and co-occurrence dynamics. This section interprets the bias trends identified by the most reliable approaches specifically Approaches 3, 8, and 9 with an emphasis on object-level associations, scene sensitivity, and dataset-conditioned variations.

Object-Level Gender Associations

Certain object categories consistently exhibited strong gender biases across all datasets and top-performing approaches. For example:

- **Male-Biased Objects**: *Baseball bat*, *skateboard*, *sports ball*, and *bicycle* were predominantly found in male-labeled images, typically embedded in outdoor, athletic, or high-action scenes. These objects received high male-bias scores under Approach 9, reinforcing their alignment with stereotypically masculine visual contexts.
- **Female-Biased Objects**: *Handbag*, *hair drier*, *umbrella*, and *cup* appeared more frequently in female-labeled images, often within domestic or personal care environments. Among these, *handbag* consistently ranked highest in female bias across all methods and datasets.

These patterns reflect socially ingrained object-gender associations and suggest that dataset composition can inadvertently encode normative gender cues. The human bias perception survey (Figure 3.14) corroborated these results, as participants labeled these same objects as “typically male” or “typically female,” confirming the interpretability and external validity of the model outputs.

Scene Context as a Bias Amplifier

Gender bias was not solely determined by object class but was significantly influenced by the surrounding scene:

- Objects like *bicycle* and *sports ball* exhibited stronger male bias when situated in outdoor or sports-related scenes.
- Conversely, *handbag* and *hair drier* showed amplified female bias in indoor, personal care, or domestic settings.

These findings validate the role of Scene Similarity Bias (SSB) in the Unified Bias Metric (UBM), emphasizing that context plays a critical role in shaping object perception and associated gender inferences.

Dataset-Driven Variability

Bias trends were not static, they adapted to the composition of the dataset:

- In the *Balanced Bias Dataset*, object scores were more symmetrically distributed across genders.
- In skewed or *Extreme Bias Datasets*, even traditionally neutral objects like *cup* or *umbrella* exhibited higher polarization, underscoring that bias attribution is influenced by contextual frequency rather than intrinsic object properties.

Approach 9 notably maintained sensitivity to these distributional shifts while preserving interpretability and score stability.

Cross-Approach Consistency

Objects identified as gender-biased by Approach 9 were consistently flagged by Approaches 3 and 8 as well. This overlap, highlighted in Figure 3.13 (Top Diverging Objects), reinforces the reliability of these methods and their capacity to detect and agree on contextually significant patterns. These observations were further supported by co-occurrence analyses (e.g., *Merged_Object_Bias_Cooccurrence.csv*), which demonstrated consistent alignment between detected bias scores and gender-labeled person presence.

The analysis confirms that gender bias in visual datasets is both object-dependent and scene-contingent. Isolated object detection is insufficient for meaningful bias detection effective models must integrate spatial context and semantic background. The Unified Bias Metric (UBM), by jointly modeling Object Influence Score (OIS) and Scene Similarity Bias (SSB), provides a lens for identifying and interpreting gendered visual cues in complex, real-world data.

Selection of the Final Unified Bias Metric

Following a rigorous evaluation of nine computational approaches for detecting contextual gender bias in image datasets, Approach 9 (SHAP + PCA → Ridge) was identified as the most effective and reliable formulation of the Unified Bias Metric (UBM). The selection was based on a multi-criteria validation framework encompassing alignment with gender misclassification patterns, coverage of object-level bias detection, score distribution stability, inter-approach correlation, human perception agreement, and model interpretability.

Across all validation dimensions, Approach 9 consistently outperformed its peers. As demonstrated in Figures 3.8 through 3.15, it exhibited:

- High alignment with real-world misclassification patterns, indicating semantic relevance of its bias predictions.
- Complete object-wise coverage (100%) across all datasets, ensuring no object class was overlooked.
- Well-balanced bias category distributions, avoiding over-reliance on “neutral” labels and maintaining sensitivity to contextual cues.
- Strong statistical correlation with other top-performing approaches, affirming its methodological stability.
- Close agreement with human-labeled bias perceptions, reinforcing the interpretability and practical trustworthiness of its outputs.

Moreover, the combination of SHAP-based interpretability, PCA-driven dimensionality insights, and Ridge regression's regularized weighting yielded a transparent and generalizable bias detection mechanism. This hybrid structure enables nuanced scoring while remaining scalable and explainable key requirements for integration into real-world fairness auditing pipelines.

In conclusion, Approach 9 is formally selected as the final and recommended configuration of the Unified Bias Metric (UBM). Its robustness, interpretability, and high empirical agreement make it well-suited for deployment in visual AI systems where ethical considerations and gender fairness are paramount.

3.3. Results And Discussion of Audio Bias Score: Bias Detection Metric for Gender Bias Detection in Audio Datasets.

3.3.1. .Introducing The Tests Used For Performance Measurement.

I. Mean Squared Error (MSE).

MSE is a fundamental metric used to measure the average of the squared differences between predicted and actual values. A lower MSE value indicates better model performance.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

- y_i = actual value
- $\hat{y}_i$ = predicted value
- n = number of data points

II. Root Mean Squared Error (RMSE).

RMSE is the square root of the MSE. It retains the same units as the target variable and provides an interpretable measure of average prediction error.

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

III. Root Mean Squared Error (RMSE).

NMSE provides a scale-invariant version of the MSE by dividing it by the variance of the actual values. It helps compare model performance across datasets with different scales.

$$NMSE = \frac{MSE}{Var(y)}$$

- $VAR(y)$ = variance of the actual values.

IV. R-squared (R^2) – Coefficient of Determination

R^2 represents the proportion of the variance in the dependent variable that is predictable from the independent variables. It ranges from 0 to 1, with values closer to 1 indicating a better fit.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - (y_i - \bar{y_i}))^2}{\sum_{i=1}^{n}(y_i - \bar{y_i})^2}$$

- $\bar{y_i}$= mean of the actual values.

V. Mean Absolute Error (MAE)

MAE measures the average magnitude of errors in predictions without considering their direction. It is more robust to outliers than MSE.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y_i}|$$

VI. Pearson's Correlation Coefficient (r)

Pearson's r measures the linear correlation between actual and predicted values. Values range between -1 (perfect negative correlation) and +1 (perfect positive correlation), with 0 indicating no linear correlation.

$$r = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y_i} - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}\sqrt{\sum_{i=1}^{n}(\hat{y_i} - \bar{\hat{y}})^2}}$$

VII. Spearman's Rank Correlation Coefficient (ρ).

Spearman's ρ is a non-parametric measure that assesses how well the relationship between two variables can be described using a monotonic function. It uses the ranks of values instead of raw values. Where, $\rho = +1$ is a perfect positive rank correlation, $\rho = -1$ is a perfect negative rank correlation and $\rho = 0$ is when no correlation is found in the ranks.

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2 - 1)}$$

- d_i= difference between the ranks of the actual and predicted values.

VIII. Kendall's Tau (τ)

Kendall's Tau is another non-parametric measure of ordinal association. It compares the number of concordant and discordant pairs. Like Spearman's ρ, the value of Kendall's τ also ranges from -1 to +1, with 0 indicating no ordinal relationship between the variables.

$$\tau = \frac{(C - D)}{\frac{1}{2}n(n-1)}$$

- C = number of concordant pairs
- D = number of discordant pairs

3.3.2. Performance Measurements.

Dataset	MSE	RMSE	MAE	NMSE	r^2
LibriSpeech (LS)	0.07173	0.26784	0.1724	0.00209	0.9979
Multilingual LibriSpeech: Italian (IT)	0.02688	0.16395	0.1149	0.00078	0.9992
Multilingual LibriSpeech: Portuguese (PT)	1.23754	1.11245	0.4622	0.03622	0.9638
Multilingual LibriSpeech: Polish (PL)	0.35646	0.59704	0.3112	0.01043	0.9896
Common voice : Hakha Chin (CNH)	0.57049	0.75530	0.5480	0.01669	0.9833
Common Voice : Chuvash (CV)	1.01335	1.00665	0.7063	0.02965	0.9703
Common Voice : Welsh (W)	0.00097	0.03125	0.0254	2.85968	0.9998
Common Voice : Kurmanji (KMJ)	0.01545	0.12433	0.0990	0.00045	0.9995
TedLium (TDL)	0.26426	0.51406	0.3697	0.00773	0.9923
The AMI Corpus (AMI)	0.00503	0.07096	0.0517	0.00014	0.99985

Table 5 MSE, RMSE, NMSE, R-Squared, MAE

Dataset	Pearson's Correlation	Spearman's Rank Correlation	Kendall Tau Rank Correlation
LibriSpeech (LS)	0.9991	0.99951	0.99534
Multilingual LibriSpeech: Italian (IT)	0.9997	0.99983	0.99767
Multilingual LibriSpeech: Portuguese (PT)	0.9848	0.98654	0.95348
Multilingual LibriSpeech: Polish (PL)	0.9954	0.99485	0.97558
Common voice : Hakha Chin (CNH)	0.9972	0.99752	0.98255
Common Voice : Chuvash (CV)	0.9962	0.99760	0.98486
Common Voice : Welsh (W)	0.9993	0.999999	0.999999
Common Voice : Kurmanji (KMJ)	0.9999	0.99999	0.99999
TedLium (TDL)	0.9970	0.99756	0.98604
The AMI Corpus (AMI)	0.9999	0.99999	0.99999

Table 6 Pearson's Correlation, Spearman's Rank, Kendall Tau Rank

I. LibriSpeech.

LibriSpeech yielded outstanding performance across all metrics. The model produced a Mean Squared Error (MSE) of 0.07173 and an R-squared value of 0.9979, indicating that nearly all variance in the target variable was explained by the model. The correlation metrics reinforce this excellent fit: the Pearson's Correlation was 0.9991, Spearman's Rank Correlation was 0.99951, and Kendall's Tau was 0.99534. These high values suggest an extremely strong linear and monotonic relationship between actual and predicted scores.

II. Multilingual LibriSpeech: Italian.

For the Italian subset of the multilingual LibriSpeech, the model's performance was exemplary. With a very low MSE of 0.02688 and an R-squared of 0.9992, the predictions closely matched the true bias scores. Furthermore, the correlation coefficients were near perfect: Pearson's at 0.9997, Spearman's at 0.99983, and Kendall's Tau at 0.99767, showing a very strong agreement in both magnitude and order of values.

III. Multilingual LibriSpeech: Portuguese.

This subset had relatively higher errors, with an MSE of 1.23754 and a lower R-squared value of 0.9638, suggesting the model struggled to generalize to Portuguese. The correlation values with Pearson's at 0.9848, Spearman's at 0.98654, and Kendall's Tau at 0.95348, while still high, were lower compared to other datasets. This could indicate variations in language structure or acoustic features that were less well captured by the model.

IV. Multilingual LibriSpeech: Polish

The Polish dataset yielded solid results, with an MSE of 0.35646 and R-squared of 0.9896. The correlation metrics were also strong: Pearson's was 0.9954, Spearman's was 0.99485, and Kendall's Tau was 0.97558. These results indicate that the model performed reliably, maintaining strong predictive ordering and correlation despite moderate error values.

V. Common Voice: Hakha Chin

For Hakha Chin, the model achieved an MSE of 0.57049 and an R-squared of 0.9833, suggesting a good fit with slightly higher residual errors. The correlation metrics—Pearson's at 0.9972, Spearman's at 0.99752, and Kendall's Tau at 0.98255—show that the predicted values preserved both the strength and rank of relationships remarkably well, despite some variance in absolute errors.

VI. Common Voice: Chuvash

Chuvash data showed one of the weaker performances with an MSE of 1.01335 and R-squared of 0.9703. Nevertheless, the correlation values remained high: Pearson's at 0.9962, Spearman's at 0.99760, and Kendall's Tau at 0.98486. This suggests that even though the model's predictions were less accurate in magnitude, the relative order and association between variables were still well preserved.

VII. Common Voice: Welsh.

Welsh data showed exceptional results with almost negligible error—MSE of 0.00097 and R-squared of 0.9998. The correlation metrics mirrored this perfection, with Pearson's Correlation at 0.9993, and both Spearman's and Kendall's Tau at 0.999999. These values indicate near-perfect linear and rank correlation, making Welsh one of the best-performing datasets in this study.

VIII. Common Voice: Kurmanji

The Kurmanji subset also performed extremely well. With an MSE of 0.01545 and R-squared of 0.9995, the model had minimal prediction error. Correlation values were similarly high—Pearson's at 0.9999, Spearman's at 0.99999, and Kendall's Tau at 0.99999—indicating an excellent fit across both value and ranking perspectives.

IX. TED-LIUM

TED-LIUM data resulted in a good model fit with an MSE of 0.26426 and R-squared of 0.9923. The correlation coefficients (Pearson's at 0.9970, Spearman's at 0.99756, and Kendall's Tau at 0.98604) suggest that the predictions were consistently aligned with the true values in both order and magnitude, despite slightly higher errors likely due to the spontaneous nature of TED talks.

X. AMI Corpus

The AMI Corpus yielded one of the most precise performances. The MSE was only 0.00503, and the R-squared was virtually perfect at 0.99985. Correlation values also reached maximum thresholds, with Pearson's at 0.9999, and both Spearman's and Kendall's Tau at 0.99999. This confirms that the model achieved nearly perfect alignment between predicted and actual bias scores in this dataset.

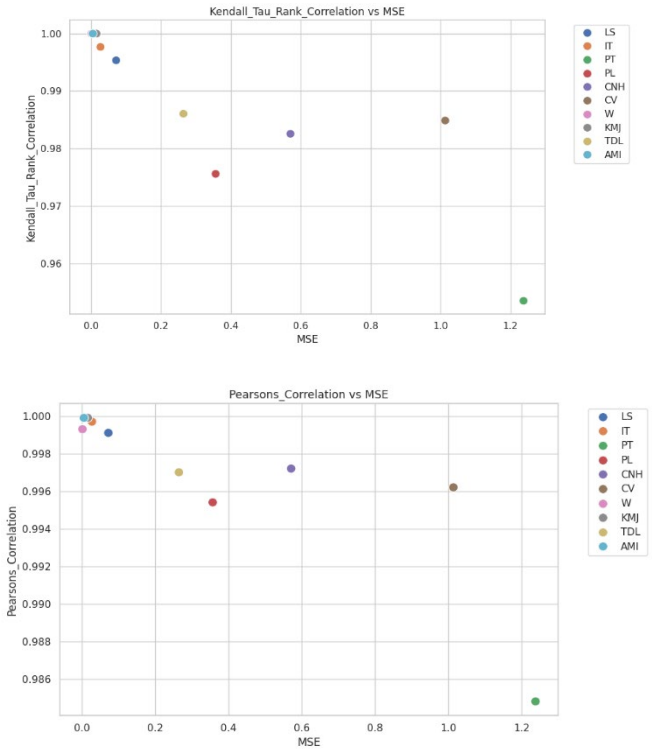

Figure 68 Kendall Tau Rank Vs MSE (top) , Pearson's Correlation Vs MSE (bottom)

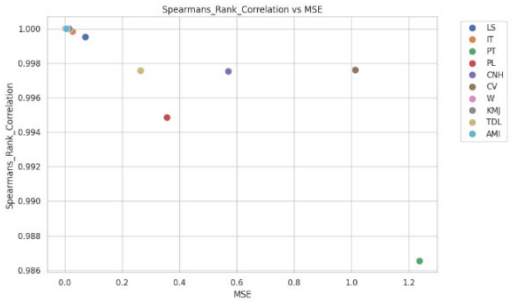

Figure 69 Spearman's Rank Vs MSE

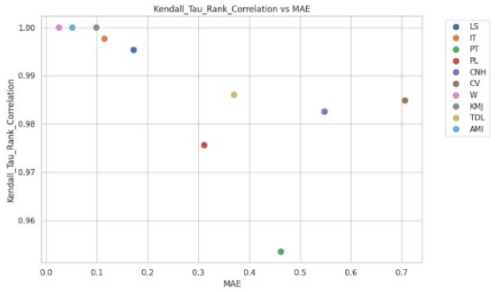

Figure 70 Kendall Tau Rank Vs MAE

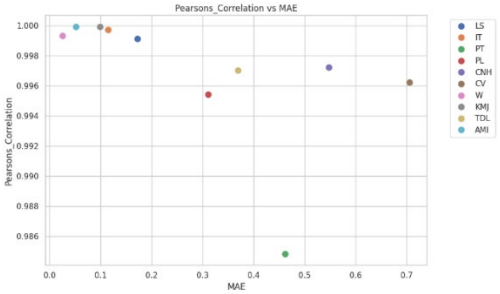

Figure 71 Pearson's Correlation Vs MAE

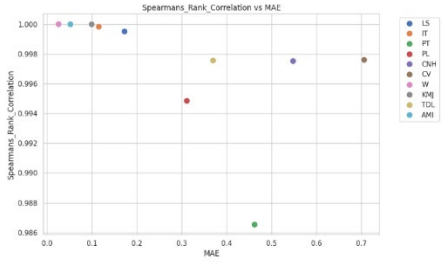

Figure 72 Spearman's Correlation Vs MAE

3.3.3. Discussion.

While the proposed bias detection metric introduces a novel and insightful approach to quantifying gender bias in audio datasets, it is important to acknowledge several limitations inherent in the current version of the metric. These limitations provide both a realistic boundary to its present capabilities and a roadmap for future enhancements and research directions.

The most significant limitation of the current metric is its exclusive focus on gender as the demographic dimension for bias analysis. While gender is a critical axis of evaluation especially in the context of fairness in AI systems, it is only one of many possible demographic attributes that can influence and skew dataset distributions and model behaviour. Important dimensions such as race, age, accent, socio-economic background, native language, and regional dialect are not captured in the current formulation of the metric. These factors, individually or in combination, can significantly affect the performance of speech-based AI systems and introduce systematic disparities that may go unnoticed if only gender is considered. For instance, the same dataset that appears balanced by gender might still demonstrate strong bias against older speakers, speakers from certain ethnic groups, or non-native speakers of a particular language. By not incorporating these dimensions, the metric potentially overlooks a wide array of intersecting and compounding biases, which could lead to incomplete or misleading interpretations of fairness in a dataset.

Another key limitation lies in the reliance on the availability of explicit gender metadata associated with the audio samples. The metric assumes that gender information is known and accurately labeled for each audio file. However, this is not always the case in real-world scenarios. Many publicly available or crowd-sourced speech datasets, such as parts of Mozilla Common Voice or TEDx corpora do not consistently provide speaker-level demographic information, or may offer gender labels based on self-identification without verification. In situations where such metadata is missing, ambiguous, or unreliable, the metric becomes either inapplicable or prone to error. While automated gender detection

techniques exist, they introduce their own risks of inaccuracy and bias, particularly when dealing with speakers whose voices deviate from typical male/female vocal patterns (e.g., children, elderly, or transgender individuals). As a result, the usability of the metric is currently confined to datasets where gender metadata is not only available.

The current implementation of the metric has been tested and validated primarily on single-speaker audio datasets. These datasets typically contain recordings where one individual speaks at a time, and the gender label is directly associated with that speaker. However, in datasets which contain meeting transcription, interview summarization, or conversational AI, audio files contain multi-speaker interactions, where multiple individuals participate in overlapping or turn-based speech. Analysing such datasets requires speaker diarization, a complex preprocessing step that segments the audio into speaker-specific portions and attributes each segment to an identified speaker. The proposed metric does not yet support or accommodate this level of complexity, and its effectiveness on multi-speaker datasets remains untested. Furthermore, errors introduced during diarization (e.g., speaker boundary errors, incorrect gender attribution) can compound the measurement inaccuracies of the bias metric.

while the proposed metric serves as a valuable starting point for understanding and quantifying gender bias in audio datasets, its current formulation has important limitations that constrain its broader applicability

Taken together, the results from all four modalities demonstrate the nuanced and multidimensional nature of demographic bias in multimodal datasets. The modality-specific bias scores provide meaningful and interpretable insights into representation gaps, while collectively laying the groundwork for the development of a unified metric. These findings reinforce the importance of domain-specific analysis in bias detection and

highlight the value of cross-modal comparisons in creating fair and inclusive AI systems. The performance of the bias equations across modalities not only validates their individual effectiveness but also informs the next steps toward building a comprehensive bias detection toolkit.

3.4. Results And Discussion of Detecting Gender Bias in Human Activity Video Datasets: A Multi-Component Visual Metric Approach

3.4.1. Bias Scores Across Activity Categories

Table 3.1 presents a comprehensive summary of gender bias scores calculated across all activity categories using three distinct metrics:

- **Mean Directional Bias Score** (Z-score standardized average of components): Indicates the direction of bias (negative = female, positive = male).
- **Bias Magnitude Score** (Sum of absolute Z-scores): Reflects the overall intensity of bias regardless of gender.
- **PCA-Weighted Bias Score** (Scaled principal component score): A composite measure that weights each metric by its contribution to variance.

These scores enable a multidimensional understanding of how gender bias manifests in different activity types. Activities such as *yoga_cat*, *yoga_bridge*, and *yoga_dancer* exhibit strong female-leaning biases across all three metrics, while *tennis_serve*, *badminton_underswing*, and *tennis_backhand* demonstrate strong male-oriented profiles.

Activity Category	Mean Directional Bias Score	Bias Magnitude Score	PCA-Weighted Bias Score
yoga_cat	-0.919092	5.979994	-0.327748
yoga_bridge	-0.770948	5.222263	-0.233092
yoga_dancer	-0.734299	4.585500	-0.256013
yoga_updog	-0.724523	5.460519	-0.162032
yoga_triangle	-0.646654	5.135346	-0.170824
gym_ride	-0.563478	5.040088	0.054926
gym_run	-0.507807	4.611012	0.116533
yoga_tree	-0.416410	4.293056	-0.000368
gym_squat	-0.226094	4.770275	-0.039488
gym_plank	-0.222724	5.389683	0.150607
gym_lift	-0.055650	3.628947	0.220343
gym_lunges	0.064973	4.297306	0.282633
backward_roll	0.095107	3.165132	0.341159
gym_push	0.185132	3.781344	0.399479
diving_jump	0.187934	2.451491	0.329341
diving_rotate	0.201408	2.396463	0.375535
gym_pull	0.293334	4.135528	0.451711
tennis_forehand	0.483019	4.008871	0.508117
golf_swing	0.502071	3.284349	0.494390
badminton_overswing	0.538724	3.831396	0.545243
backflip	0.546839	3.270509	0.559145
badminton_serve	0.603412	4.643824	0.599659
tennis_backhand	0.665701	4.115206	0.618860
badminton_underswing	0.675600	4.481038	0.627961
tennis_serve	0.744424	4.377540	0.645459

Table 7 : Activity-Wise Bias Scores Across Metrics

3.4.2. Gender Bias Trends in Activity Classes

To explore how gender bias manifests across different physical activity categories, the combined bias scores were analyzed at the category level. This section interprets the results using three key metrics derived from standardized bias components: **Directional Bias Score**, **Bias Magnitude Score**, and **PCA-Weighted Bias Score**.

Directional Bias Score

Figure 3.1 presents the average **directional bias** for each activity class, computed as the mean of standardized bias components. A positive score indicates a male-leaning bias, whereas a negative score signals a female-leaning tendency.

Notably, activities such as *tennis_serve*, *badminton_underswing*, and *tennis_backhand* exhibit strong male bias, while categories such as *yoga_cat*, *yoga_bridge*, and *yoga_dancer* reveal pronounced female bias. This dichotomy reflects the influence of traditional gender participation patterns in sports versus wellness-focused domains like yoga.

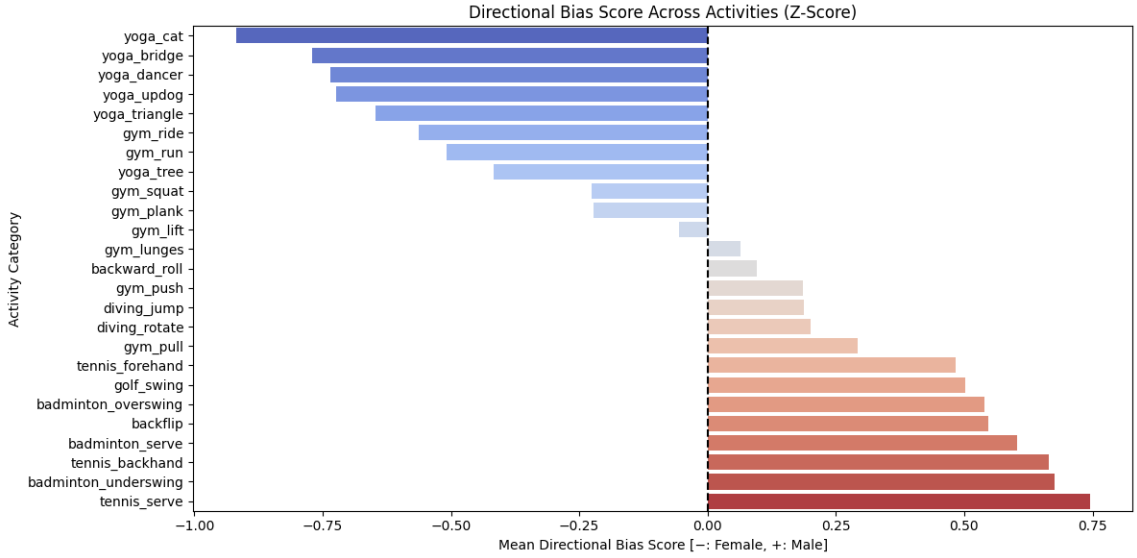

Figure 73 :Average directional bias score by activity category. Positive scores indicate male bias; negative scores indicate female bias.

Bias Magnitude Score

To assess the overall strength of gender bias regardless of direction, the sum of absolute standardized component values was computed (Figure 3.2). A higher magnitude implies stronger deviation from gender neutrality, even if the activity is evenly distributed between genders.

Categories such as *yoga_cat*, *yoga_bridge*, and *gym_plank* top the list in bias magnitude, revealing not just directional skew but significant visual or motion-based disparities between genders. Conversely, mid-range activities like *gym_push* and *diving_jump* show moderate bias intensities.

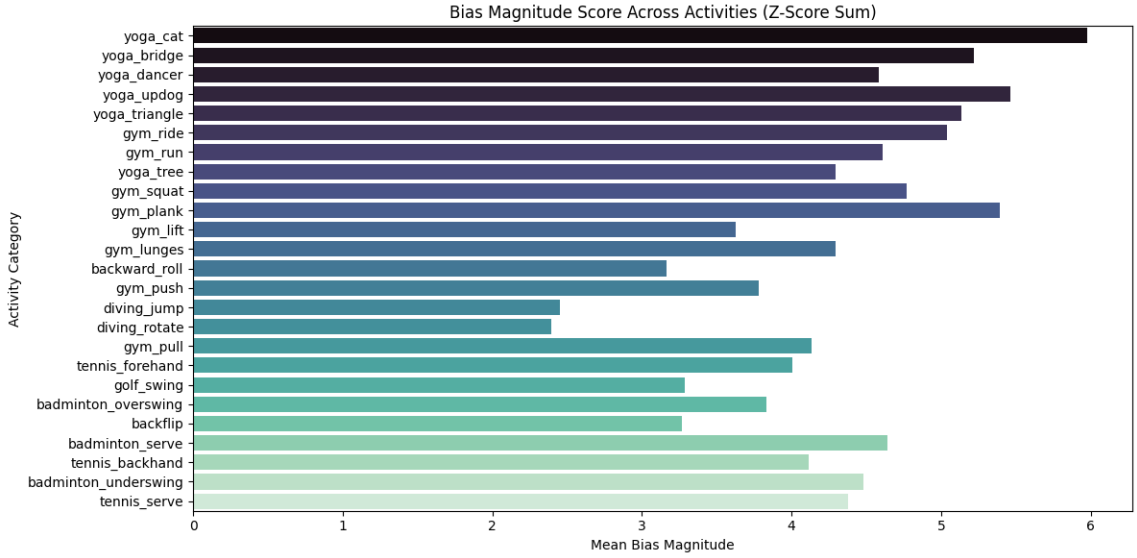

Figure 74 :Average bias magnitude by activity category. Higher values indicate stronger visual representation bias across components.

PCA-Weighted Bias Score

The PCA-weighted score captures the most explanatory dimension of variation in bias using Principal Component Analysis (Figure 3.3). Here, dimensional weights emphasize metrics contributing most to global bias variability.

Results show that male-dominant sports such as *tennis_serve*, *badminton_underswing*, and *backflip* have high positive PCA-weighted scores, while female-associated categories like *yoga_cat*, *yoga_dancer*, and *yoga_bridge* are positioned strongly in the negative direction. Interestingly, some categories such as *gym squat* show mild directional bias but a stronger PCA-weighted leaning, suggesting nuanced gender distinctions captured through motion and embedding-based features.

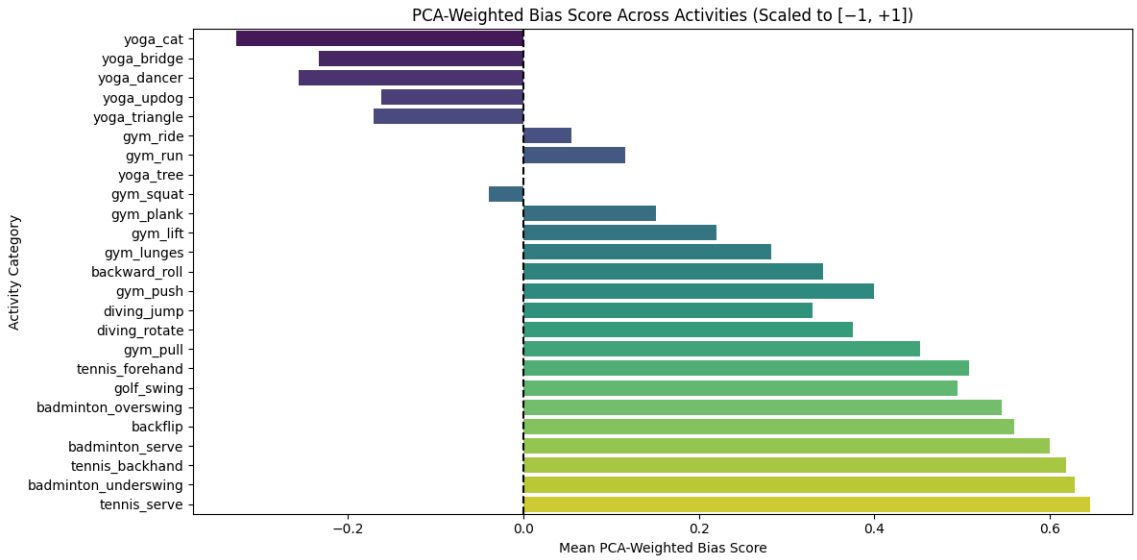

Figure 75 :PCA-weighted bias scores, highlighting categories where deeper visual features skew toward a particular gender

Interpretation Summary

- **Yoga categories** consistently show strong female bias across all three views, aligning with demographic trends in practice participation.
- **Racket and combat sports** lean strongly toward male bias, reflecting visual, semantic, and motion differences tied to performance styles.
- **PCA-weighted scores** provide sharper differentiation in classes with mixed gender representation, highlighting the utility of multivariate aggregation.

3.4.3. Gender Composition and Bias Correlation

To understand the interplay between gender distribution and visual bias, activity categories were examined in terms of their gender representation and corresponding bias

metrics. Table 3.2 presents the top ten **male-dominant** and **female-dominant** categories based on the number of videos labeled for each gender.

In the female-dominant set, categories such as *yoga_dancer*, *yoga_cat*, and *yoga_bridge* demonstrated extreme gender skew, with *yoga_dancer* containing no male samples. These categories also ranked among the most female-biased in both directional and PCA-weighted metrics, indicating strong visual, spatial, and motion characteristics more typical in female-labeled samples. The alignment between representation imbalance and bias scores is particularly evident in yoga-based classes, which also recorded the highest bias magnitude scores, suggesting consistently distinctive features across components.

Conversely, male-dominant categories such as *tennis_serve*, *badminton_underswing*, and *backflip* were associated with high positive values in both directional **and** PCA-weighted bias scores. These results reflect not only demographic skew but also substantial visual and motion-specific emphasis toward male-labeled samples. The strong correlation in these categories indicates that participation imbalance amplifies measurable bias in both component and composite metrics.

However, not all categories with strong bias exhibited large gender disparities. For example, *gym_plank* and *gym_push* displayed moderate gender counts but still produced noticeable directional and PCA-weighted biases. These cases suggest that content-level factors such as pose movement, camera framing, or screen prominence contribute independently to visual bias beyond gender frequency alone.

🔵 Top Male-Dominant Categories			🔴 Top Female-Dominant Categories		
Category	**M**	**F**	**Category**	**M**	**F**
backflip	18	2	yoga_dancer	0	20
tennis_serve	17	3	yoga_cat	2	18
tennis_backhand	17	3	yoga_bridge	2	18

badminton_underswing	16	4	gym_squat	3	17
badminton_serve	16	4	yoga_triangle	3	17
badminton_overswing	15	5	yoga_updog	4	16
golf_swing	15	5	yoga_tree	5	15
gym_pull	14	6	gym_ride	6	14
tennis_forehand	14	6	gym_run	7	13
gym_push	13	7	gym_plank	9	11

Table 8 :Most gender-dominant activity categories based on video counts (Top 10 per gender

3.4.4. Interpretation and Significance

The integration of component-level bias metrics provides deeper insights into which features, such as spatial prominence, motion dynamics, or semantic embeddings, contribute to perceived gender biases in video-based datasets.

- Bias is feature-dependent:

  The use of multiple metrics revealed that some activities (e.g., *gym_plank*, *gym_push*) displayed relatively balanced gender representation but still exhibited bias in motion or pose-centric components. This suggests that bias can emerge from stylistic framing, not just frequency.

- Directionality ≠ Magnitude:

  While some activities may lean slightly toward one gender, the magnitude metric emphasizes whether this tilt is visually and semantically strong. Thus, a small directional score can still correspond to strong bias if variation across metrics is high.

- PCA-weighted bias generalizes variance:

  The use of PCA assigns weights to each metric based on how much they contribute to inter-video variance. Activities with high PCA scores (e.g., *tennis_serve*,

badminton_serve) show not only directional tilt but also feature-distinct clustering, suggesting stronger model-learnable bias.

- Correlation with representation:

Bias scores correlate with gender distribution, but not always linearly. Some categories with relatively balanced gender count still exhibit high composite bias, implying that qualitative visual encoding (like pose expressivity or centrality) influences perception and algorithmic weighting more than raw counts.

3.4.5. Discussion

This study presented a multi-component framework to detect and quantify gender bias in human activity video datasets, using five distinct visual and motion-based metrics: Size Bias, Centering Bias, Screen Time Bias, Embedding Bias, and Motion Bias. These metrics were standardized and combined into three composite scores—directional, magnitude, and PCA-weighted—to provide a robust and interpretable view of bias distribution.

Findings revealed clear patterns of gender bias across activity categories. Activities such as *yoga_dancer* and *yoga_cat* consistently exhibited strong female-leaning scores, while *tennis_serve* and *badminton_serve* showed pronounced male-leaning bias. Notably, some classes demonstrated high bias despite having balanced gender representation, suggesting that visual framing, motion patterns, and pose articulation significantly contribute to representational skew. This indicates that dataset fairness cannot be assessed solely by subject frequency but requires deeper feature-level analysis.

The contributions of this research include a novel bias metric design, an effective aggregation strategy, and an empirical analysis pipeline built on widely-used vision tools (YOLOv8, MediaPipe, SlowFast). The framework offers a scalable and replicable method for dataset auditing, suitable for integrating into larger fairness evaluation workflows. Looking forward, extending the methodology to support non-binary labels, conducting perceptual validation, and testing across diverse datasets will enhance its inclusivity and

generalizability. This work lays the groundwork for more equitable vision systems by enabling deeper diagnostics of bias in human activity recognition.

4. Conclusion

This thesis presented a multimodal framework for detecting and quantifying gender bias in AI datasets across four modalities: text, image, audio, and video. Each modality-specific metric was independently designed, validated, and integrated to address the limitations of existing bias detection methods.

In the **text modality**, the Context-Aware Bias Metric (CABM) was introduced to evaluate gender bias in contextual word embeddings using three features: cosine similarity, PMI, and contextual bias from sentence embeddings. Among the tested weighting strategies, the SHAP + Random Forest method proved most effective yielding balanced, interpretable, and statistically robust bias scores. CABM revealed clear gender-stereotypical patterns and demonstrated the superiority of dynamic, context-sensitive approaches over static metrics like WEAT and MAC.

In the **image modality**, the Unified Bias Metric (UBM) combined Object Influence Score (OIS) and Scene Similarity Bias (SSB) across nine computational approaches. The final selected configuration Approach 9 (SHAP + PCA + Ridge Regression) showed high alignment with misclassification trends, full object-level coverage, and strong agreement with human perception. It offered the best balance between interpretability, statistical significance, and methodological stability, validating its suitability for fairness auditing in vision-based AI systems.

For **audio datasets**, a bias metric was developed based on acoustic features such as pitch, energy, and duration. The approach effectively captured representation disparities between male and female voices. High statistical correlations and low prediction errors across diverse corpora (e.g., AMI, TED-LIUM) confirmed the reliability of the bias score,

although the current formulation was limited to single-speaker datasets and binary gender representation

In the **video modality**, a multi-component visual metric framework was proposed to quantify gender bias in human activity recognition datasets. The framework evaluated five distinct dimensions, Size Bias, Centering Bias, Screen Time Bias, Embedding Bias, and Motion Bias, using deep learning models including YOLOv8, MediaPipe, and SlowFast. These metrics were standardized and aggregated into directional, magnitude, and PCA-weighted scores to assess both the strength and orientation of bias. Empirical results revealed that visual and motion-based gender disparities persisted across several activity categories, including sports and fitness contexts. Importantly, some activities exhibited strong representational bias even in the absence of gender imbalance, suggesting that visual framing and movement patterns play a critical role. This modality-specific metric enabled scalable bias auditing and added interpretability to fairness evaluation in video-based machine learning systems.

Across all modalities, the proposed metrics validated through statistical testing and human-based evaluation offer interpretable, scalable, and modular solutions for fairness auditing. This multimodal framework lays the groundwork for a comprehensive bias detection toolkit that can be extended to other protected attributes and adapted for real-world deployment.

5. References

[1] Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean, "Efficient Estimation of Word Representations in Vector Space," 2013.

[2] Jeffrey Pennington, Richard Socher, Christopher D. Manning, "GloVe: Global Vectors for Word Representation," 2014.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," 2019.

[4] Caliskan A, Bryson JJ, Narayanan A., "Semantics Derived Automatically from Language Corpora Contain Human-like Biases.," 2017.

[5] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, Kai-Wei Chang, "Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods," 2018.

[6] Krizhevsky, Alex & Sutskever, Ilya & Hinton, Geoffrey., "ImageNet Classification with Deep Convolutional Neural Networks. Neural Information Processing Systems," 2012.

[7] Joseph Redmon, Santosh Divvala, Ross Girshick, Ali Farhadi, "Joseph Redmon, Santosh Divvala, Ross Girshick, Ali Farhadi," 2016.

[8] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, Ilya Sutskever, "Learning Transferable Visual Models From Natural Language Supervision.," 2021.

[9] Joy Buolamwini, Timnit Gebru, "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification.," 2018.

[10] A. Wang, A. Narayanan, and O. Russakovsky, "REVISE: A Tool for Measuring and Mitigating Bias in Visual Datasets," European Conference on Computer Vision, 2020.

[11] Rachel K. E. Bellamy, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, Seema Nagar, Karthikeyan Natesan Ramamurthy, John Richards, Diptikalyan Saha, Prasa, "AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias," 2018.

[12] Hilde Weerts, Miroslav Dudík, Richard Edgar, Adrin Jalali, Roman Lutz, Michael Madaio, "Fairlearn: A Toolkit for Assessing and Improving Fairness in AI.," 2023.

[13] Greenberg, Craig & Mason, Lisa & Sadjadi, Seyed & Reynolds, Douglas, "Two Decades of Speaker Recognition Evaluation at the National Institute of Standards and Technology," 2019.

[14] Md Abu Sayed, Maliha Tayaba, MD Tanvir Islam, Md Eyasin Ul Islam Pavel, Md Tuhin Mia, Eftekhar Hossain Ayon, Nur Nob, Bishnu Padh Ghosh, "Parkinson's Disease Detection through Vocal Biomarkers and Advanced Machine Learning Algorithms," 2023.

[15] Ko, Hyunwoong & Kwon, Sukbong., "Optimization of Voice Biomarkers to Predict Alzheimer’s Disease," 2023.

[16] Jahangir, Rashid & Wah, Teh & Nweke, Henry & Mujtaba, Ghulam & Al-Garadi, Mohammad & Ihsan, Ali., "Speaker Identification through Artificial Intelligence Techniques: A Comprehensive Review and Research Challenges," 2021.

[17] Wiebke Hutiri, Tanvina Patel, Aaron Yi Ding, Odette Scharenborg, "As Biased as You Measure: Methodological Pitfalls of Bias Evaluations in Speaker Verification Research," 2024.

[18] Nicole Meister, Dora Zhao, Angelina Wang, Vikram V. Ramaswamy, Ruth Fong, Olga Russakovsky, "Gender Artifacts in Visual Datasets," 2023.

[19] Ahmed Sabir, Lluís Padró, "Women Wearing Lipstick: Measuring the Bias Between an Object and Its Related Gender," 2023.

[20] Rs, Ramprasaath & Cogswell, Michael & Das, Abhishek & Vedantam, Ramakrishna & Parikh, Devi & Batra, Dhruv., "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," 2020.

[21] Haofan Wang, Zifan Wang, Mengnan Du, Fan Yang, Zijian Zhang, Sirui Ding, Piotr Mardziel, Xia Hu, "Score-CAM: Visual Explanations via Gradient-Free Localization," 2020.

[22] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, Kai-Wei Chang., "Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods," arXiv, 2018.

[23] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Association for Computational Linguistics*, 2019.

[24] A. S. P. K. R. F. F. H. a. B. H. S. Schröder, "Evaluating metrics for bias in word embeddings," arXiv, 2021.

[25] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, Piotr Dollár, "Microsoft COCO: Common Objects in Context," 2015.

[26] Jocher, G., Chaurasia, A., Stoken, A., Borovec, J., NanoCode012, Kwon, Y, "YOLOv8 by Ultralytics: Real-Time Object Detection and Image Segmentation," 2023.

[27] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, Ross Girshick, "Segment Anything," 2023.

[28] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, Hengshuang Zhao, "Depth Anything: Unleashing the Power of Large-Scale Unlabeled Data," 2014.

[29] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, Ilya Sutskever, "Learning Transferable Visual Models From Natural Language Supervision," 2021.

[30] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva and A. Torralba, "Places: A 10 Million Image Database for Scene Recognition," IEEE Transactions on Pattern Analysis and Machine Intelligence (Volume: 40, Issue: 6, 01 June 2018), 2018.

[31] M. Shabbir, A. Hussain and M. M. Khan, "Age and Gender Estimation Through Speech: A Comparison of Various Techniques," in *2023 18th International Conference on Emerging Technologies (ICET)*, Peshawar, Pakistan, 2023.

[32] A. Ghosal, C. Pathak, P. Singh and S. Dutta, "Voice-Based Gender Identification Using Co-occurrence-Based Features," *Computational Intelligence in Pattern Recognition,* pp. 947-956, 2020.

[33] E. Priya, S. J. Priyadharshini, P. S. Reshma and S. Sashaank , "Temporal and spectral features based gender recognition from audio signals," in *2022 International Conference on Communication, Computing and Internet of Things (IC3IoT)*, Chennai, India, 2022.

[34] Y. Ali, E. Noorsal, N. F. Mokhtar, S. Z. Md Saad, M. H. Abdullah and L. C. Chin, "Speech-based gender recognition using linear prediction and mel-frequency

cepstral coefficients," *Indonesian Journal of Electrical Engineering and Computer Science,* vol. 28, no. 2, p. 753, 2022.

[35] S. Fahmeeda, M. A. Ayan, M. Shamsuddin and A. Amreen, "Voice Based Gender Recognition Using Deep Learning," *International Journal of Innovative Research & Growth,* vol. 3, pp. 649-654, 2022.

[36] S. Srivastava, H. Sharma and D. Garg, "Comparative Study of Machine Learning Algorithms for Voice based Gender Identification," in *2022 International Conference on Edge Computing and Applications (ICECAA)*, Tamilnadu, India, 2022.

[37] Z. Zhang , R. Li and K. Chen, "Speaker Gender Recognition Based on Semi-Supervised Learning," in *2024 3rd International Conference on Computing, Communication, Perception and Quantum Technology (CCPQT)*, Zhuhai, China, 2024.

[38] I. A. Destina and E. Hartati, "An Analysis of intonation pattern in the pre-service english teacher talks," *FRASA : English Education and Literature Journal,* vol. 3, no. 2, pp. 64-71, 2022.

[39] S. Ekeruke, "Stress Patterns and Intonations among the Annang Language Speaking Students of Faculty of Arts, Akwa Ibom State University," *Journal of Cmmunication and Culture,* vol. 12, no. 3, pp. 163-172, 2024.

[40] B. Ludusan, M. Heldner and M. Wlodarczak, "Exploring the role of formant frequencies in the classification of phonation type," in *Proceedings of 20th International Congress of Phonetic Sciences*, Prague, Czech Republic, 2023.

[41] A. Bailey and M. D. Plumbley, "Gender Bias in Depression Detection Using Audio Features," in *Proceedings of the 29th European Signal Processing Conference (EUSIPCO)*, Dublin, Ireland, 2021.

[42] R. Ardila, M. Branson, K. Davis, M. Kohler, J. Meyer, M. Henretty, R. Morais, L. Saunders, F. Tyers and G. Weber, "Common Voice: A Massively-Multilingual Speech Corpus," in *Proceedings of the Twelfth Language Resources and Evaluation Conference*, Marseille, France, 2020.

[43] V. Panayotov, G. Chen, D. Povey and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, South Brisbane, QLD, Australia, 2015.

[44] V. Pratap , Q. Xu, A. Sriram, G. Synnaeve and R. Collobert, "MLS: A Large-Scale Multilingual Dataset for Speech Research," in *Interspeech 2020*, Shanghai, China, 2020.

[45] A. Rousseau, P. Deléglise and Y. Estève, "TED-LIUM: an Automatic Speech Recognition dedicated corpus," in *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, 2012.

[46] I. McCowan, J. Carletta, W. Kraaij, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, . T. Hain, J. Kadlec, . V. Karaiskos, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, W. Post, D. Reidsma and . P. Wellner, "The AMI meeting corpus," in *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC)*, Marrakech, Morocco, 2008.

[47] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *ICML'23: Proceedings of the 40th International Conference on Machine Learning*, Honolulu Hawaii USA, 2023.

[48] C.-h. W. H.-r. Y. Y.-W. T. C.-K. T. Jihoon Chung, "HAA500: Human-Centric Atomic Action Dataset with Curated Videos," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, Montreal, Canada, 2021.

[49] Z. Y. T. W. e. a. Haoyi Qiu, "Gender Biases in Automatic Evaluation Metrics: A Case Study on Image Captioning," in *arXiv.org*, 2023.

[50] I. T. G. E. e. a. Dimitrios Kastaniotis, "Gait-based Gender Recognition Using Pose Information for Real Time Applications," in *International Conference on Digital Signal Processing (DSP)*, Santorini, Greece, 2013.

[51] K. J. S. K. S. A. S. R. U. K. Dwivedi, "An Overview of Moving Object Detection Using YOLO Deep Learning Models," in *IEEE International Conference* , Dehradun, India, 2024. , 2024.

[52] C. e. a. Lugaresi, ""MediaPipe: A Framework for Building Perception Pipelines," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* , Long Beach, USA, 2019.

[53] H. F. J. M. K. H. C. Feichtenhofer, "SlowFast Networks for Video Recognition," in *IEEE International Conference on Computer Vision (ICCV)*, Seoul, South Korea, 2019.

[54] Haodong Duan, Yue Zhao, Kai Chen, Yuanjun Xiong, Dahua Lin, "Mitigating Representation Bias in Action Recognition: Algorithms and Benchmarks," 2022.

[55] Tianlu Wang, Jieyu Zhao, Mark Yatskar, Kai-Wei Chang, Vicente Ordonez, "Balanced Datasets Are Not Enough: Estimating and Mitigating Gender Bias in Deep Image Representations," 2018.

[56] P. -S. Tan, S. Rajanala, A. Pal, S. -M. Leong, R. C. . -W. Phan and H. Fang Ong, "Causally Uncovering Bias in Video Micro-Expression Recognition," 2024.

[57] C. e. a. Lugaresi, "MediaPipe: A Framework for Building Perception Pipelines," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Long Beach, USA, 2019.

[58] H. F. J. M. K. H. Christoph Feichtenhofer, "SlowFast Networks for Video Recognition," in *IEEE International Conference on Computer Vision (ICCV)*, Seoul, South Korea, 2019.