

**UNIVERSITY OF THE WITWATERSRAND,
SCHOOL OF ELECTRICAL AND INFORMATION
ENGINEERING**

ELEN4000A/4011A: Electrical Engineering Design II

Topic 7: AI for All - Building an Online Platform to Eliminate Bias

Natan Grayman - 2344104

28 August 2023

AI Bias Mitigation Tool

Abstract: In the realm of artificial intelligence (AI), the surge of transformative advancements is accompanied by the vexing challenge of inherent bias within AI models, which poses the potential for perpetuating discrimination. To address this concern, this project aims to craft an innovative online bias mitigation tool grounded in a contextual and human-centric approach. By synergizing the importance of comprehending the overarching contextual framework with a methodological stance towards identified sources of bias within the machine learning pipeline, a systematic step-by-step UI design solution is meticulously forged. This exploration encompasses the facets of historical bias, representation bias, measurement bias, evaluation bias, aggregation bias, and deployment bias, fortified by a suite of cutting-edge open source explainability, interpretability, and fairness metrics. The web application serves as a conduit for presenting these results through a model card-inspired curated report, channelling actionable insights for countering the negative impact of bias. Furthermore, the architecture of the web application capitalizes on robust AWS technology, ensuring a synergy of security, scalability, privacy, crowdsourcing, maintainability, cost-effectiveness, and environmental sustainability. This synthesis culminates in a versatile bias mitigation tool poised to address bias across a spectrum of societal domains.

Keywords: Machine Learning, Bias, Fairness Metrics, Explainability, Interpretability.

Table of Contents

1. Introduction..... 1

2. Background..... 1

 2.1 AI Model Governance..... 1

 2.1.1 Human-Centred Design..... 2

 2.1.2 Model Cards..... 2

 2.2 Bias Taxonomy 2

3. Bias Mitigation Design 3

 3.1 Historical Bias..... 3

 3.1.1 Explainability..... 3

 3.1.2 Reweighing Algorithm..... 4

 3.1.3 Fairness Metrics 5

 3.2 Representation Bias 6

 3.2.1 Adversarial Debiasing..... 6

 3.2.2 Domain Adaptation 6

 3.3 Measurement Bias..... 6

 3.4 Evaluation Bias 7

 3.5 Aggregation Bias..... 7

 3.6 Deployment Bias..... 8

 3.7 Curated Report..... 8

4. Detailed System Architecture Design 9

 4.1 Frontend Design..... 9

 4.2 Scalability, Fault Tolerance and Security..... 9

 4.3 Backend Data Handling 11

 4.4 Serverless Data Processing and Integration..... 11

 4.5 Machine Learning Integration..... 12

5. Discussion..... 13

 5.1.1 Cost Analysis 13

 5.1.2 Environmental Sustainability and Carbon Footprint..... 14

6. Conclusion 14

References..... 15

1. Introduction

The rapid integration of artificial intelligence (AI) into a multitude of decision-making processes across the digital landscape has ushered in an era of transformative advancements. However, alongside these opportunities, there exists a pressing need to comprehend the intricate mechanisms by which AI systems make decisions and to critically evaluate their outcomes.

Consequently, this design project central aim is to design and establish an innovative online platform dedicated to the assessment and mitigation of inherent bias in AI models. This pioneering platform will offer an invaluable resource for both researchers and practitioners, facilitating the uploading of AI models and datasets for systematic bias evaluation. Leveraging cutting-edge techniques from the realm of machine learning, including state-of-the-art fairness metrics, interpretability, explainability methods and open-source algorithms, the platform’s design is aimed to unravel and quantify bias within AI models. Moreover, it will extend beyond mere diagnosis, providing expert-driven insights and recommendations to alleviate bias and advance fairness and transparency in AI systems.

This report encompasses a multifaceted exploration of bias within AI models, ranging from its conceptual foundation to the practical application of mitigation strategies and system design considerations. The subsequent sections delve into these pivotal dimensions, commencing with the background section addressing AI model governance and bias categorization, followed by the step-by-step bias mitigation design. This is succeeded by an exploration of the comprehensive architecture of the online platform and an ensuing discussion on costs and environmental sustainability.

2. Background

This background section lays the foundation for comprehending the proposed design solution and its associated challenges and intricacies. The design presented in this report is centred around tabular data constraints and is further constrained to apply to the model agnostic approach, treating the model as an input-output with varying weights. The section will initiate by highlighting the significance of AI model governance and drawing inspiration from pertinent examples that shape the proposed solution. Subsequently, the bias taxonomy will be examined explaining how bias demands a deep understanding of its nuanced expressions and its capacity to creep into the machine learning (ML) pipeline, ultimately impacting downstream decision outcomes [1].

2.1 AI Model Governance

AI model governance embodies the framework through which organizations regulate access, enforce policies, and monitor activities pertaining to machine learning models and their outcomes [2]. Effective model governance encompasses diverse practices such as setting stringent access controls for deployed models, maintaining version histories, creating comprehensive documentation, ongoing monitoring of models, and aligning machine learning practices with established organizational IT policies and user needs.

In tandem with these governance efforts, the AI community has witnessed the emergence of several influential initiatives that guide the ethical considerations and responsible development of AI systems. Notably, the IEEE P7003 standard provides methodologies to address and eliminate negative biases during algorithm creation, encompassing aspects of safety, transparency, accountability, and minimizing bias [3]. Likewise, interdisciplinary forums, such as the ACM Conference on Fairness, Accountability, and Transparency (ACM FAccT), serve as catalysts for fostering progressive dialogues on crucial subjects including fairness, algorithmic accountability, and the ethical ramifications of AI [4]. These initiatives

underscore the growing emphasis on aligning AI technologies with ethical guidelines and societal values. These initiatives underscore the growing emphasis to instil accountability, traceability, and responsible usage of AI models, thereby proactively mitigating potential risks and ensuring alignment with regulatory guidelines and privacy mandates [5]. Informed by these principles, two prominent instances exemplifying AI model governance have come to the forefront: human-centred design and model cards.

2.1.1 Human-Centred Design

Human-centricity is considered a central aspect in the development and governance of artificial intelligence (AI), aiming to adapt the concept of human-centred design (HCD) to the public governance context of AI [6]. This endeavour has given rise to the term "Human-Centric AI" (HCAI) [6]. Importantly, HCD highlights the integration of human decision-making processes across all phases of the design journey [6]. HCD represents a research approach that leverages diverse viewpoints and multidisciplinary teams to gain insights into the contextual nuances and cognitive processes of both data and individuals interacting with AI systems [6]. Employing an iterative methodology, HCD highlights a feedback-driven approach and complements, rather than substitutes, human intelligence in systems [7]. Consequently, integrating the principles of HCD into this web application will inherently cultivate a design that prioritizes fairness, transparency, and accountability, effectively catering to the diverse spectrum of user groups.

2.1.2 Model Cards

Model cards play a crucial role in standardizing documentation procedures aimed at reporting the performance characteristics of ML models [8] [9]. Serving as a compact summary, model cards provide insights into the intended applications of AI capabilities and potential impacts on various user groups [8] [9]. By providing a comprehensive overview of an AI model's behaviour, potential biases, limitations, and intended usage, model cards offer stakeholders a customizable documentation tool that facilitates more informed and transparent decision-making [8] [9].

2.2 Bias Taxonomy

A common misconception regarding bias in machine learning algorithms is the notion that undesirable behaviours within ML systems solely arise due to "biased data" [1]. While it is undeniably true that flawed data plays a substantial role in the propagation of bias within ML systems, attributing all undesirable and prejudiced outcomes solely to "data bias" oversimplifies the issue [1]. In actuality, the machine learning pipeline is complex and multifaceted, whereby bias has the potential to infiltrate various stages throughout the model development process [10].

Exploring the realm of bias within machine learning is a complex undertaking, especially for practitioners new to the topic [11]. Determining the sources of bias poses a formidable challenge, further compounded by the extensive usage of bias terminology and the presence of ambiguous vocabulary [12] [13]. For instance, terms like 'algorithmic' bias or 'systemic' bias can hold varied interpretations based on their origins, leaving room for ambiguity. This plethora of different sources of bias necessitate precise definitions, as well as a clear framework and categorization of bias within the context of the machine learning pipeline [1] [11].

In the literature, several distinct organizational principles for creating such a taxonomy exist. Within the field, a diversity of perspectives exists regarding the classification of bias and its mitigation techniques. For example, one comprehensive study aiming to provide a broad multidisciplinary overview of the area of bias in AI systems, categorizes bias with the distinction among issues stemming from data generation, data collection, or institutional bias [11]. Whereas, in [12] group types of biases are constructed on how they interact with either the data, the algorithm, or the user.

Nevertheless, the taxonomy that provides the most suitable and structured framework for effectively addressing the multifaceted challenge of bias in machine learning (ML) has been identified in [1]. This taxonomy, depicted in Figure 1, serves as the chosen framework for understanding the sources of harm throughout the ML lifecycle.

Importantly, each stage depicted is not mutually exclusive, characterizing them as distinct sources of bias simplifies the process and enhances comprehension [1] [10]. This approach acknowledges the interconnections and interdependencies of various stages while providing a coherent structure and methodological framework for addressing bias comprehensively.

In essence, the chosen taxonomy fosters a holistic, strategic, and adaptable approach to bias mitigation throughout the machine learning pipeline. By providing a well-structured, model-agnostic, and prospective-oriented framework, it enables the creation of a customized, focused, and versatile structure for designing the bias mitigation techniques within the online application.

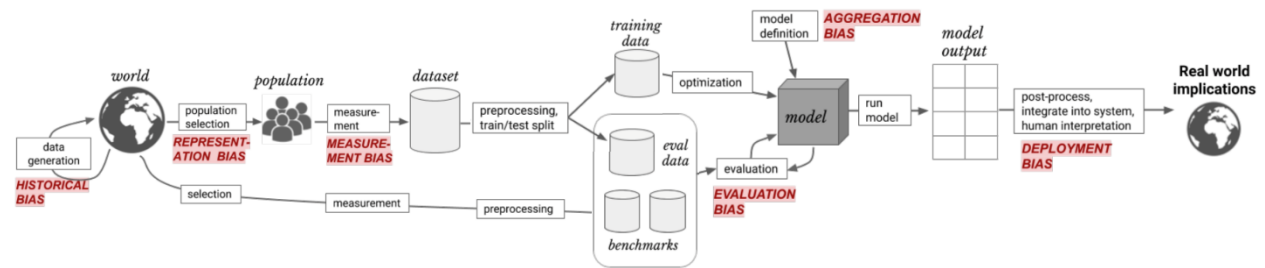


Figure 1: Depiction of the framework of bias sources throughout the ML pipeline, adapted from [1].

3. Bias Mitigation Design

This section thoroughly explores every facet of the bias taxonomy, delineating essential steps and effective mitigation techniques, while also highlighting the significance of the context-specific curated report.

3.1 Historical Bias

Historical bias, a critical aspect of bias in machine learning, arises when the data originates from a flawed state of the world [14]. Understanding the existence of bias within the data generation process is crucial, as it has the potential to perpetuate inequalities and generate unjust predictions downstream [1]. Even if a model accurately reflects historical or current realities, it can still lead to harm within a particular population [1]. Therefore, ignoring this bias can result in predictions that further entrench social disparities, impede societal progress, and diminish trust in automated systems [14].

3.1.1 Explainability

The first step in the process of mitigating bias for all sources of harm downstream involves the utilization of explainability [15]. Explainability, within the context of machine learning, refers to the capacity to elucidate the decisions and reasoning behind the predictions made by AI models using human-readable terms [11] [15]. This comprehension not only aids in diagnosing bias but also guides the selection and implementation of appropriate mitigation techniques that align with the dataset's intricacies [14] [15]. Thus, utilizing explainability metrics initially will help identify the demographic attributes in the dataset—protected attributes such as sex, gender, race, age, ethnicity, etc. By comparing the distribution of these attributes to their real-world proportions, potential underrepresentation can be identified, enabling the mitigation of historical bias.

Upon delving into the existing literature on explainability concepts, several prominent techniques have garnered attention for their effectiveness in shedding light on AI model decision-making. Firstly, permutation importance stands out as a heuristic method for gauging feature importance by assessing how a feature's shuffling impacts model performance [16]. This approach aids in ranking features based on their influence; however, it falls short in providing a detailed understanding of the specific nature and direction of each feature's effect [15] [16]. Furthermore, partial dependence plots (PDP) offer a valuable visualization tool to showcase how altering a single feature's value influences model predictions while keeping other features constant [15] [16]. Although insightful for grasping the impact of individual attributes, these plots do not capture potential interactions among multiple features [16].

Shapley Additive explanations (SHAP) values introduce a game-theoretic approach to unravel the output of machine learning models [16]. This technique stands out for its ability to provide a comprehensive explanation of feature contributions, effectively unveiling the intricate interplay of features within the model's predictions [15] [16].

By aggregating SHAP values, a generative summary plots that visually depict the cumulative effect of different features on model predictions can be produced. This holistic view helps understand not only the magnitude of each feature's influence but also how their combined effects shape predictions [15] [16]. By aggregating SHAP values, a generative summary plot that visually depicts the cumulative effect of different features on model predictions can be produced [15]. This holistic view helps explain not only the magnitude of each feature's influence but also how their combined effects shape predictions. It specifically facilitates the assessment of historical bias by allowing for an evaluation of protected attributes, such as the reinforcement of stereotypes, associated with a particular group. Nevertheless, the selected explainability metric of SHAP values will serve as foundational reference and utilization point for addressing all sources of bias. Its values enable a deeper comprehension of the intricate relationships between features and predictions, rendering it an indispensable tool for the web application's mission of fostering transparency and fairness within AI models.

3.1.2 Reweighting Algorithm

Exploring advanced techniques for analysing and mitigating historical bias unveils a spectrum of strategies aimed at rectifying disparities. Techniques such as Generative Adversarial Networks (GANs) present an innovative approach by training a GAN to generate data points that mimic the existing dataset, introducing more diversity [17]. However, within the context of mitigating historical bias with flawed data, the introduction of synthetic data through GANs could inadvertently exacerbate existing biases or even introduce new biases, thereby posing significant challenges in achieving fairness and equitable outcomes [17].

Resampling, encompassing both oversampling and undersampling, provides options to rectify imbalance by either duplicating underrepresented group instances or reducing those from overrepresented ones [18]. However, while resampling techniques like oversampling and under-sampling can address class distribution imbalances, they come with limitations [19]. Oversampling may inflate certain instances, compromising dataset integrity, while undersampling could remove informative data, limiting model learning [19]. Both strategies also heighten the risk of overfitting, weakening model performance and potentially overfitting to flawed historical data [1] [19].

The re-weighting algorithm emerges as a particularly compelling choice to help mitigate historical bias due to its ability to strategically reweigh the specific protected attributes [20]. By assigning varied weights to different groups, particularly elevating those for underrepresented categories, the algorithm ensures that the original data's characteristics and features are preserved without compromising the overall dataset structure

[16] [20]. Moreover, IBM's AI Fairness 360 (AI360) open-source library offers a reweighing algorithm. The transparency and accessibility of this open-source platform are reinforced by IBM's reputable standing in the industry. The library's comprehensive and well-documented features streamline the process of integration and implementation [21]. It offers python code that seamlessly integrates with the Jupyter notebook provided by Amazon SageMaker, accompanied by guiding materials, documentation, and notebooks [21]. These resources facilitate the translation of algorithmic research from the laboratory to practical domain applications [21]. Thus, the versatile reweighing algorithm, along with its open-source resources, serves as a valuable tool for mitigating historical bias and quantifying fairness metrics.

3.1.3 Fairness Metrics

Fairness metrics encompass various criteria to define equity in machine learning (ML) models [22]. In the context of this design, four essential fairness metrics will be explored specifically demographic parity, equal opportunity, equal accuracy, and group unaware.

Demographic parity evaluates a model's fairness by comparing the distribution of selected individuals to the distribution of applicants across different groups [22] [23]. The model is deemed fair if the proportions of selected individuals from each group match the group membership percentages of the applicants. This metric seeks to ensure equitable representation among selected groups [22]. Equal opportunity, on the other hand, fairness aims to maintain the same true positive rate (TPR) or sensitivity for each group [22]. This fairness metric guarantees that the proportion of individuals correctly selected by the model ("positives") is consistent across all groups [22] [23].

Equal accuracy fairness focuses on maintaining uniform accuracy rates across all groups [23]. This metric ensures that the percentage of correct classifications, including both approvals and denials, is the same for all groups [23]. If the model achieves a specific accuracy level for one group, it should attain the same level for other groups as well [22]. Alternatively, group unaware fairness involves removing group membership information from the dataset [23]. By eliminating factors such as gender, this approach aims to create a model that remains impartial to various demographic groups [23].

Each of these metrics offers a distinct perspective on fairness, addressing different aspects of bias and disparity within machine learning models. By exploring these metrics, the design aims to identify effective strategies to enhance fairness and mitigate biases in the AI system.

Nevertheless, the Impossibility Theorem of Machine Fairness asserts that it's statistically impossible for a single example to fulfil them all at the same time [24]. This raises the ambiguity of which fairness criterion should take precedence and which metric to choose [24]. Furthermore, in selecting the appropriate fairness metric, the interplay between parity versus preference and treatment versus impact plays a pivotal role [22]. Parity versus preference prompts the query of whether fairness entails achieving demographic parity or meeting individual preferences [22]. Treatment versus impact probes whether fairness should be upheld during the treatment process or in the resultant impacts and outcomes [22].

Evidently, the selection of a fairness criterion is a complex endeavour that demands thorough deliberation involving all stakeholders. Its determination is highly subject to the precise context and project objectives. Hence, within the user interface (UI) designed, users are guided to provide in-depth context by answering key questions. This structured approach involves progressing step-by-step through the questions relevant to the specific source of bias. Through this process, an overview is established, enabling the experts of the online application to assess the uploaded datasets and models within the user's contextual framework. Subsequently, a curated report is created for the customer, factoring in the context and stakeholder information, and consequently proposing the suitable mitigation technique and fairness metric.

3.2 Representation Bias

Representation bias arises when the development dataset inadequately represents certain segments of the population, leading to poor representation for specific subgroups [1]. This bias can stem from several factors, including the target population's misalignment with the development dataset, underrepresented groups within the population, or biased sampling methods [1]. For instance, data that is representative of Johannesburg may not accurately represent the population of Cape Town when used for analysis; likewise, data depicting Johannesburg's demographics from three decades ago may not accurately reflect the current population of the city.

3.2.1 Adversarial Debiasing

Adversarial debiasing, is an in-processing technique that harmonizes two objectives: maximizing prediction accuracy and minimizing the adversary's ability to deduce the protected attribute from predictions [25]. Contained within the IBM AI360 open-source library, adversarial debiasing provides a versatile approach to addressing representation bias, supported by research that highlights its effectiveness compared to alternative debiasing methods [26]. Firstly, it aligns the model by training it to predict the protected attribute while minimizing its influence on the main prediction task, thus reducing group discrimination stemming from disparate data distributions [26]. Moreover, it addresses underrepresentation by incorporating an adversarial network that encourages the model to disregard protected attributes, ensuring fair treatment across all groups during prediction [1] [26].

3.2.2 Domain Adaptation

Considering the intricacies of representation bias, domain adaptation will also be recommended as an integral aspect of the web application design, primarily for its capability to mitigate biased sampling methodologies [1] [27]. Domain Adaptation aims to align the distribution of data between the source domain (where the model is trained) and the target domain (where the model will be deployed), effectively reducing the bias introduced by differences in data distribution [27] [28]. Consequently, an expert well-versed in adversarial bias could leverage domain adaptation strategies, recommending the incorporation of data from related domains where underrepresented groups are better represented. The collaborative interplay between expert-guided adversarial debiasing and domain adaptation specifically targets representation bias, providing a versatile approach to achieve heightened fairness, bolstered generalization, and enhanced predictive accuracy.

3.3 Measurement Bias

Measurement bias, often referred to as detection bias, is a form of bias that emerges when the selection, collection, or computation data varies across groups [1] [23]. It encompasses any non-random or systematic error that arises during data collection, affecting the accuracy and reliability of the collected data [1]. In essence, measurement bias pertains to the discrepancies between the chosen measurements or proxies, such as features or labels, and the underlying abstract constructs they aim to represent [1].

To address measurement bias, various mitigation strategies can be utilized. Notably, data auditing and quality assurance emerge as impactful approaches due to their ability to systematically identify and rectify biases and inaccuracies within the dataset [29] [30]. Data auditing entails a meticulous assessment of data quality, diversity, and completeness, aiming to uncover and correct potential biases or inaccuracies [29]. On the other hand, quality assurance entails validating data proxies for labels or features to uphold their accuracy and consistency [30].

To implement these mitigation strategies, it is proposed to leverage the inherent advantage of hand-labelled data in identifying measurement bias [31]. Integrating human-centred hand-labelled data proves effective in mitigating measurement bias as it enables navigation through the multifaceted intricacies of the contextual landscape, ensuring a more robust and equitable model [31].

Initially, the data auditing and quality assurance will be carried out by the expert team of the online application. However, if the workload becomes too substantial for the small team, the proposed design solution is to utilize Amazon Mechanical Turk (MTurk) [32]. Amazon MTurk is a platform for crowdsourcing that simplifies the process of outsourcing tasks and assignments to a widely distributed workforce [32]. It proves to be an applicable choice for outsourcing data quality and auditing, thanks to its ability to scale with a diverse workforce, cost-effective hiring reductions, and seamless integration with Amazon SageMaker, all of which collectively accelerate the necessary workload effectively [32]. To ensure user privacy, a standardized policy will be implemented before outsourcing data through Amazon Mechanical Turk. This will include anonymizing data by removing personally identifiable information (PII) and refraining from collecting such information from workers, aligning with MTurk's Acceptable Use Policy [32].

3.4 Evaluation Bias

Evaluation bias refers to the distortion of model performance metrics due to improper comparison methods or flawed benchmark datasets [1] [23]. It arises from the desire to quantitatively compare different models and make general statements about their quality [1] [23]. However, such comparisons can lead to invalid conclusions, especially when benchmark datasets are affected by historical, representational, or measurement bias [1].

To mitigate these sources of harm within the benchmark dataset, the proposed design solution allows for the benchmark data to be uploaded and can thus undergo any of the mitigation techniques previously discussed, thereby evaluating its fairness and applicability. Furthermore, the expert-driven team will conduct research, investigating any published flaws.

3.5 Aggregation Bias

Aggregation bias materializes when groups are improperly merged, leading to a model that lacks optimal performance for any specific group or solely excels for the predominant group [1] [23]. This bias stems from the assumption that the mapping between inputs and labels remains consistent across all data subsets, which often doesn't align with reality [1]. Diverse backgrounds, cultures, or norms can render the same variable with disparate implications across groups [1]. The result might be a model that is suboptimal for all groups or overfits to the predominant population [1].

LIME (Local Interpretable Model-agnostic Explanations) stands out as a robust mitigation strategy for tackling aggregation bias, offering a nuanced understanding of model behaviour that transcends the limitations imposed by aggregated data [33]. Its effectiveness lies in its ability to zoom in on individual instances within the dataset, meticulously dissecting the model's decision-making process [33]. This granular examination enables LIME to spotlight the unique interactions between features and predictions for distinct subgroups and instances, a capability vital for exposing the disparities introduced by aggregation bias [33].

Furthermore, LIME seamlessly integrates into the Amazon SageMaker environment, facilitated by its inclusion in IBM's AI360 suite and its dedicated Python library [21]. Thus, by leveraging LIME's interpretability capabilities as well as its alignment with the model-agnostic constraints, its inclusion

effectively enhances the bias mitigation tools' functionalities and offers an enhanced method of scrutinizing the harm caused by aggregation bias.

3.6 Deployment Bias

Deployment bias, the final facet of bias in the taxonomy under consideration, arises when there is a disparity between the intended problem-solving scope of a model and its practical application [1]. This incongruity often stems from constructing a model with the assumption of complete autonomy, while in practice, the system operates within intricate sociotechnical frameworks guided by institutional dynamics and human decision-making processes [34]. Thus, comprehending the contextual landscape of model deployment and user interactions holds paramount importance.

To effectively address deployment bias, the platform will leverage the power of crowdsourcing as a potent strategy for user validation, continuous monitoring, and feedback. Crowdsourcing involves engaging users to collectively evaluate and rectify any inaccuracies within the bias detection algorithms' predictions. To facilitate this process, the chosen crowdsourcing capabilities will be implemented through Vanilla Forums.

Vanilla Forums provides diverse capabilities to elevate user engagement and privacy [35]. The web application will harness its open-source nature and tap into the ability to create private forums exclusively accessible to invited members through a link to Vanilla Forums [36]. Empowering expert moderators from the web application with complete control over the forum's elements, including appearance, functionality, and discussions, aligns with ethical practices [36]. This approach nurtures a safe environment conducive to user input, monitoring and validation.

Furthermore, this forum will prove to be a valuable channel for collecting feedback. Users will be encouraged to report instances where they observe the bias detection results to be inaccurate or biased, drawing from their real-world experiences and interactions with the model. They can provide detailed insights into their particular use cases and model interactions, thereby enriching the dataset that informs the refinement and enhancement of bias detection algorithms. Ultimately, this aligns with the platform's overarching objective of a human-centred approach that champions fairness in AI models.

3.7 Curated Report

As discussed in section 3.1.3, the curated report is generated by the expert team behind the online application, drawing inspiration from the concept of model cards to instil a human-centric contextual approach into addressing bias [23] [37]. To illustrate this design concept, a prototype of the curated report has been developed, as depicted in Figure 6. This example provides a contextual overview of a test concerning the malignancy status of breast cancer tumours, encompassing sections like model details, intended use case, model type, and an analysis of both training and evaluation data.

Moreover, this prototype includes customized visualizations to elucidate the data analysis process, featuring the previously discussed SHAP values and a confusion matrix illustrating the contrast between true and predicted outcomes for two distinct racial groups [15] [23]. Leveraging this prototype information, the expert team offers actionable insights through fairness metrics, highlighting the areas where the data exhibits flaws. In essence, this prototype of the curated report effectively showcases its intended purpose and significance, forming a foundation for subsequent refinement and expansion. When coupled with the expert driven array of algorithmic tools and techniques, it establishes a methodological, holistic, and adaptable framework for designing a bias mitigation tool.

4. Detailed System Architecture Design

The frontend and the cloud infrastructure design has been meticulously tailored to align with the bias mitigation step-by-step process, ensuring security, maintainability, redundancy, and scalability throughout the system. The comprehensive cloud infrastructure overview is depicted in Figure 7. Amazon's AWS emerged as the obvious selection for cloud infrastructure, owing to its remarkable array of products and the necessary tools that seamlessly integrate with its ecosystem. This section delves into the frontend design choice and the intricacies of the cloud infrastructure design image, offering insights into the purpose and functionality of each component, while highlighting how they symbiotically contribute to an effective solution for the system.

4.1 Frontend Design

Framer was chosen as the optimal tool for web design and prototyping, due to its distinct advantages over conventional frontend development methods, particularly during the initial project stages [38]. Framer's canvas-based approach enables designs to be directly translated into interactive prototypes, enhancing the efficiency, and reducing inconsistencies between design and execution [38]. This choice enables swift deployment, reduces unnecessary development costs, and streamlines the software development lifecycle, allowing a stronger focus on the core bias mitigation tool.

Moreover, Framer ensures top-notch quality, scalability, maintainability, and creative freedom for collaborative efforts [38]. It delivers a responsive website with Search Engine Optimization (SEO) capabilities, offers customization for business pricing and domains, and supports seamless integration into the entire marketing stack [38] [39]. The tool's real-time collaboration features facilitate teamwork between designers and developers, promoting optimized code through joint efforts [38]. Thus, the maintainability of the platform will be enhanced, providing longevity to the platform's functionality, adaptability to future updates and improvements, and overall sustainability in the face of changing technologies and user needs [39].

Ultimately, pairing these features with Framer's ability for immediate web application deployments, with its seamless integration with Amazon Web Services (AWS) [38], solidifies Framer's value as an excellent choice for the web application and was thus utilised to prototype the frontend design.

4.2 Scalability, Fault Tolerance and Security

The client's journey begins by accessing the platform through the URL. The URL name is registered with the Amazon Route 53. Amazon Route 53, a highly available and scalable Domain Name System (DNS), which seamlessly resolves the web address to an IP address, ensures efficient redirection of users to the appropriate internet gateway hosted on AWS [40]. This will be configured to direct users to the Canonical Name (CName) of the created AWS Cloud Application Load Balancer [41]. Therefore, when the client enters the URL, they will be directed to the Application Load Balancer.

The Application Load Balancer will be equipped with rules to assess the operational status of high availability firewalls located across different availability zones within the Virtual Private Cloud (VPC), ensuring availability and functionality [41]. VPC serves as a segregated network environment within the AWS cloud infrastructure [42]. Within a secure VPC, data and resources are isolated, providing enhanced security and control over network configurations, access policies, and communication between components [42]. Furthermore, by spreading the VPC over multiple availability zones physical geographical redundancy is introduced within the cloud architecture.

Auto Scaling, an AWS service, adjusts instance numbers based on demand changes, ensuring smooth scalability [33]. A Network Virtual Appliance (NVA), is a software-defined networking solution, offering flexibility, scalability, and efficient traffic control. Within each availability zone, a public subnet is established, isolated by the Auto Scaling NVA, which acts as a boundary between the public and private subnets. This design is akin to a Demilitarized Zone (DMZ), ensuring secure separation between internet-facing components in the public subnet and sensitive backend resources residing in the private subnet.

The Load Balancer will distribute incoming traffic through the Auto Scaling Network Virtual Appliance (NVA) firewall, when these firewalls are functioning properly and active [41] [43]. It's superior to traditional firewalls and seamlessly integrates into AWS for enhanced network security and traffic management within the cloud environment [43] [44].

The process of user authentication within the application, in tandem with the integration of Amazon Identity and Access Management (IAM) with PostgreSQL, constitutes a pivotal aspect of establishing secure access to digital resources [45] [46]. This fusion with AWS IAM offers distinct security advantages, granting secure control over users' interactions with AWS services and resources [45]. Specifically, in the context of PostgreSQL, IAM's credentials, encompassing an Access Key ID and Secret Access Key, play a pivotal role in ensuring secure application authentication [45]. This integration harnesses IAM database authentication, utilizing an authentication token instead of a password, thereby eliminating the necessity for password hashes [45]. Moreover, PostgreSQL stands out as an exceptional choice, primarily owing to its attributes as an open-source relational database, known for its simplicity in setup, operation, and scalability within cloud environments [46]. Furthermore, its cost-effectiveness, flexibility in hardware capacity, and complemented high maintainability provided by IAM underscore its appropriateness for bolstering the application's security [46].

After passing through the Auto Scaling Network Virtual Appliance (NVA) firewall, the traffic proceeds to the internal elastic load balancer. This elastic load balancer plays a crucial role by assessing the health and availability of both the Web Application Firewalls (WAFs) and the Web Frontends [47] residing in a purpose-built Front End VPC. This assessment is made possible by the Transit Gateway, an AWS networking service that enables seamless communication between multiple VPCs and replaces the need for intricate peering relationships, functioning as a highly scalable cloud router [48].

The Web Application Firewalls (WAFs) are strategically positioned to provide an enhanced layer of security [49], effectively defending against OWASP (Open Web Application Security Project) attacks, including SQL injections and cross-site scripting [50]. Positioned to sanitize incoming web requests at Layer 7 of the OSI model, which represents the application layer, they serve as a robust defence mechanism against such attacks [49]. By analysing and filtering incoming requests, the WAFs contribute to ensuring the integrity and security of the web application.

Utilizing Amazon EC2 C5 instances for the frontend provides a dynamic and cost-effective solution tailored for running compute-intensive tasks, particularly those involving machine learning applications [51] [52]. The c5.xlarge instance, a notable member of the compute-optimized family, is equipped with 4 vCPUs, 8.0 GiB of memory, and remarkable hardware components [52]. The hardware CPU's advanced architecture ensures efficient handling of complex computations, rendering it exceptionally effective for powering machine learning algorithms [52]. Additionally, the instance's robust capacity allows for up to 10 Gbps of bandwidth, a vital attribute that seamlessly supports the demands of the web application, especially when dealing with high data throughput. [52] Thus, the c5.xlarge instance strikes an optimal balance between computational performance, affordability, and high bandwidth, making it a strategic choice for driving the frontend of the application.

In essence, this multi-layered designed architecture aligns with the platform's emphasis on scalability and fault tolerance, enabling seamless adaptation to varying user loads while maintaining redundancy and failover mechanisms to ensure high availability.

4.3 Backend Data Handling

Upon user upload of databases, models, and key information, the platform's data management strategy involves utilizing Amazon S3 buckets within the framework of the Amazon Data Lake Formation [53]. An Amazon S3 bucket, short for Amazon Simple Storage Service, serves as a robust and scalable object storage solution provided by AWS [54]. This strategic implementation aligns with the platform's commitment to efficient, private, and secure data handling.

S3 buckets, organized hierarchically with a flat namespace, facilitate schema-less data storage [54]. This schema flexibility eliminates the need for a predefined structure, allowing diverse data formats to be stored without constraints and upload errors [54]. Nevertheless, to optimize data handling further, the platform recommends the utilization of the open-source HDF5 (Hierarchical Data Format version 5) data structure type.

HDF5 offers an array of benefits that make it an excellent fit for data management, including an efficient data compression for storage, chunking for optimized storage and retrieval of large data, support for parallel input/output to enhance distributed computing performance, a wide range of supported data types, compatibility with the h5py package within Amazon SageMaker Python SDK and efficiency in managing large training datasets for machine learning applications [55] [56]. Furthermore, this preferred data structure format efficiently encapsulates both the model's architecture and weights, ensuring consistency between the user's locally trained model and the model analysed in the bias mitigation process [55]. Therefore, by incorporating the uploaded .h5 weights, the platform seamlessly integrates the user's expertise and efforts into the bias mitigation process, resulting in a collaborative and impactful approach to addressing bias in AI models.

Leveraging S3 buckets within the Data Lake Formation, takes advantage of advanced features like versioning, data encryption, and fine-grained access controls [53] [54]. Firstly, data lake formation's Data Catalog [57] enhances privacy measures. By cataloguing metadata about the data assets stored in the data lake, including each model and respective dataset the platform creates a comprehensive view of the data landscape [57]. Secondly, within the Data Lake Formation, Access controls can be applied to regulate who can access specific datasets, columns, or rows, ensuring that sensitive information is restricted to authorized individuals or groups [58].

Additionally, the platform employs Amazon CloudTrail, a service that tracks user activity within the AWS cloud environment [59]. CloudTrail records all user interactions with the data, enabling detailed auditing and monitoring of data access by providing an extensive log of actions taken on the data, including information about who accessed which data and when [59]. This level of transparency and accountability is particularly valuable for maintaining privacy and adhering to data protection regulations. It ensures that any access to sensitive data is thoroughly documented and traceable, thereby contributing to a robust compliance framework and reinforcing the platform's commitment to privacy and ethical data handling practices.

4.4 Serverless Data Processing and Integration

Amazon API Gateway, a powerful service provided by AWS, simplifies the process of developing, deploying, and managing APIs on a scalable level [60]. It offers developers the tools to effortlessly create,

publish, maintain, monitor, and secure APIs, regardless of their scale [60]. One of the standout features of API Gateway is its compatibility with serverless workloads and web applications [60]. Therefore, it is a suitable choice for effectively directing incoming requests from the user interface to the underlying AWS Lambda functions.

AWS Lambda functions as the core of serverless data processing within the designed cloud platform, allowing code execution without the need for server provisioning [61]. By eliminating the need for manual server provisioning, AWS Lambda enables python code execution in a highly efficient and resource-optimized manner [61] [62]. This service encapsulates discrete tasks within Lambda functions, enabling modular, manageable, and rapid software lifecycle and data processing operations [61] [62].

Moreover, with AWS Lambda's pay-as-you-go model and flexible resource allocation, users can efficiently allocate memory to functions, allowing AWS Lambda to automatically distribute proportional CPU power, network bandwidth, and disk input/output (I/O) resources [62]. Therefore, leveraging Lambda functions, the platform can seamlessly process user-uploaded datasets and models, abstracting the complexities of infrastructure management and offering a streamlined and effective solution within the cloud environment.

In the context of the machine learning pipeline, AWS Step Functions offer a powerful tool to manage and automate the various stages of data ingestion efficiently [63]. This orchestration ensures a seamless flow of tasks, allowing for easy integration of subsequent services like AWS Glue, which comprises of AWS DataBrew and AWS Glue ETL [64].

AWS Glue DataBrew is a versatile tool that facilitates direct exploration and experimentation with data from diverse sources, notably including data lake formations integrated with S3 buckets [64]. The effectiveness of AWS Glue DataBrew lies in its rich set of features, providing access to more than 250 prebuilt transformations [64]. These transformations encompass a wide range of actions, from filtering anomalies to standardizing formats and rectifying invalid values [64]. This capability ensures that any uploaded datasets, models, or forum answers undergo thorough and consistent data cleaning processes, contributing to the overall quality and reliability of the platform's machine learning pipeline.

AWS Glue ETL (Extract, Transform, Load) further enhances the backend data pipeline by providing a serverless, automated and scalable solution for data extraction, transformation, and loading tasks [64]. It's a suitable choice due to the need for accuracy and consistency in handling and processing large datasets [64]. Thus, this process is essential for ensuring that data is appropriately structured for downstream tasks, optimizing the integration with Amazon Athena, and setting the stage for insightful bias exploration.

4.5 Machine Learning Integration

Amazon Athena is an interactive query service that enables seamless data analysis directly within Amazon S3 using standard SQL [65]. Its serverless architecture eliminates the need for infrastructure setup or management, and charges are incurred only based on the queries executed [66]. Athena streamlines data exploration by facilitating rapid, interactive queries on processed data stored within the data lake's processed Amazon S3 bucket, removing the need for data formatting and infrastructure management [66].

Amazon SageMaker serves as a fully managed machine learning service that facilitates the data scientists and developers to build, train and analyse machine learning models quickly and efficiently [67]. Integrated with Amazon Athena using the PyAthena library, SageMaker introduces an integrated Jupyter authoring notebook instance, eliminating the need for server management and providing effortless access to data sources for exploration and analysis [66] [67].

Furthermore, Amazon SageMaker offers a range of benefits that contribute to the robustness and effectiveness of the cloud platform. Firstly, SageMaker employs Role-Based Access Control (RBAC), which ensures that only authorized developers with specific access levels can interact with the platform [66]. This meticulous access control, following the principle of least privilege (POLP), safeguards sensitive user information, reinforcing security and privacy protocols. Secondly, SageMaker provides a suite of common machine learning algorithms that are meticulously optimized to efficiently handle substantial amounts of data within a distributed environment. This enables data scientists and developers to analyse and process large-scale datasets effectively [67]. Lastly, SageMaker's native support for bring-your-own-algorithms and frameworks offers an exceptional solution for leveraging open-source algorithms and expert-driven techniques [67]. This capability empowers the platform to effectively analyse, detect, and mitigate potential sources of bias within the client's machine learning pipeline, further enhancing the fairness and reliability of the AI models.

5. Discussion

In this section the costs and environmental considerations will be discussed.

5.1.1 Cost Analysis

The expenses associated with team salaries for the development of the web application amount to R1,449,000, as detailed in table 1. To incentivise the recruitment and retention of skilled and dedicated employees, an increase of 50% has been applied to the average salary within each category. Additionally, the commitment to attracting top talent includes provisions such as offering equity to senior experts, providing opportunities for professional growth through workshops and certifications, and granting research autonomy to allow researchers to explore projects within the company's domain.

Moreover, for projecting the anticipated usage costs of the AWS infrastructure designed, AWS's pricing calculator was employed, illustrated in figure 8 [68]. The calculated total amounts to an estimated monthly cost of \$18,498.85, equivalent to approximately R344,510 at a conversion rate of 18.62. Capital investment will be required to cover the initial development costs. However, the AWS infrastructure provides a versatile pay-as-you-go model, ensuring accessibility for small, medium, and large companies. Consequently, revenue needs to be calculated on an ad hoc basis, ensuring adequate funds to accommodate the specific dataset and model requirements of each company.

Table 1: Salary Costs for Development Team

Expense	Cost (R/month)	Information
Software Developers	585 000	Team of 3 senior software developers, proficient in full stack, AWS cloud infrastructure. R130 000 average p.m. [69].
Data Scientists	225 000	Team of 2 Data Scientists proficient in AWS infrastructure, python, and Amazon SageMaker. R75 000 average p.m. [70].
Social Scientists	96 000	Team of 2 Social Scientists with AI ethics research. R32000 average p.m. [71].
Marketing manager	150 000	1 Marketing Manager to develop marketing plan. R100 000 average [72] p.m.

Expert ML researcher	395 000	3 Expert ML researchers strong background in advanced algorithms, data analysis, and model evaluation R87500 average p.m. [73].
Total	R1 449 000	

5.1.2 Environmental Sustainability and Carbon Footprint

In 2019, Amazon committed to enhancing sustainability and minimizing carbon emissions [74]. As part of this commitment, they co-founded The Climate Pledge, with a resolute aim to attain carbon neutrality by 2040 [75]. In alignment with this goal, AWS effectively employs its innovative Customer Carbon Footprint Tool [74]. This invaluable resource empowers users to gain comprehensive insights into the environmental impact of their applications, concisely summarizing carbon emissions and associated savings [74].

Accordingly, to convey this critical information to users, a dedicated section within the curated report has been designed to display these statistics from the Carbon Footprint Tool. These statistics specifically pertain to the user's AI models and datasets used on the online application's cloud infrastructure and use metric tons of carbon dioxide equivalent (MTCO₂e) to illustrate the carbon footprint [74]. As exemplified in Figure 6 of the prototype customized report, pivotal environmental metrics include total carbon emissions and a breakdown showcasing AWS's carbon reduction accomplishments through renewable energy initiatives and energy-efficient hardware [74]. Ultimately, leveraging AWS's cloud infrastructure integrated with sustainable energy initiatives and resource-efficient hardware underscores the bias mitigation tool's commitment to a responsible and eco-conscious design solution.

6. Conclusion

In conclusion, while the evolution of AI offers numerous prospects, the intrinsic bias issue within these models presents substantial risks of propagating discrimination. In response to these dynamics, this project aims to design an innovative online bias mitigation tool, employing a contextual and human-centric approach. By amalgamating the significance of comprehending the overarching contextual framework with a holistic and methodological approach to the identified sources of bias within the machine learning pipeline, a systematic step-by-step UI design solution was formulated. Specifically, the exploration encompassed historical bias, representation bias, measurement bias, evaluation bias, aggregation bias, and deployment bias, accompanied by a suite of cutting-edge open source explainability, interpretability, and fairness metrics to counter their adverse effects. The web application effectively showcases these outcomes through a model card inspired curated report, presenting actionable insights to mitigate the detrimental impact of bias. Moreover, the web application's system architecture leverages robust AWS technology, ensuring security, scalability, privacy, crowdsourcing, maintainability, cost-effectiveness, and environmental sustainability. This amalgamation results in a versatile bias mitigation tool, aptly equipped to address bias across diverse societal domains.

References

- [1] H. Suresh and J. Gutttag, “A Framework for Understanding Sources of Harm throughout,” Association for Computing Machinery, New York, NY, USA, 2021.
- [2] DataRobot, “What is Model Governance?,” Boston, Massachusetts, 2020.
- [3] I. S. Association, “Algorithmic Bias Considerations,” IEEE, 2017.
- [4] A. f. C. Machinery, “ACM Conference on Fairness, Accountability, and Transparency (ACM FAccT),” ACM, 2023.
- [5] A. Dafoe, “AI Governance: A Research Agenda,” Centre for the Governance of AI, Future of Humanity Institute, University of Oxford, Oxford, 2017.
- [6] A. Sigfrids, J. Leikas, H. Salo-Pontinen and E. Koskimies, “Human-centricity in AI governance: A systemic approach,” Faculty of Management and Business, Administrative Sciences, Tampere University, Tampere, Finland, Jyväskylä, Finland, 2023.
- [7] IBM, “What is human-centered AI?,” Armonk, New York, 2002.
- [8] M. Mitchell, S. Wu, A. Zaldivar, P. Barnes and L. Vasserman, “Model Cards for Model Reporting,” Cornell University, Ithaca, NY, 2018.
- [9] SHANKAR, VAR; Cook, Alexis;, “Model Cards,” Google, 2023. [Online]. Available: <https://www.kaggle.com/code/var0101/model-cards/tutorial>.
- [10] ALEXIS COOK, “Identifying Bias in AI,” 2023. [Online]. Available: <https://www.kaggle.com/code/alexisbcook/identifying-bias-in-ai/tutorial>.
- [11] E. Ntoutsi, P. Fafalios, . U. Gadiraju, . V. Iosifidis, . W. Nejdl, . M.-E. Vidal, S. Ruggieri, F. Turini, . S. Papadopoulos and . E. K. Krasanakis, “Bias in data-driven artificial intelligence systems—An introductory survey,” WIREs Data Mining and Knowledge Discovery, 2020.
- [12] N. Mehrabi, F. Morstatter, N. Saxena, . K. Lerman and . A. Galstyan, “A Survey on Bias and Fairness in Machine Learning,” Cornell University, Ithaca, NY, 2019.
- [13] Google, “Fairness: Types of Bias,” 2023. [Online]. Available: <https://developers.google.com/machine-learning/crash-course/fairness/types-of-bias>. [Accessed 28 August 2023].

- [14] Reva Schwartz , Apostol Vassilev , Kristen Greene , Lori Perine , Andrew Burt , Patrick Hall, “Towards a Standard for Identifying and Managing Bias in Artificial Intelligence,” NIST Special Publication 1270.
- [15] Kaggle, “Machine Learning Explainability,” 2023. [Online]. Available: <https://www.kaggle.com/learn/machine-learning-explainability>. [Accessed 28 August 2023].
- [16] V. P. ,. S. K. Pantelis Linardatos, “Explainable AI: A Review of Machine Learning Interpretability Methods,” Department of Mathematics, University of Patras, 26504 Patras, Greece, Patras, 2020.
- [17] J. P.-A. M. M. ,. X. ,. D. W.-F. S. O. A. C. Y. B. Ian Goodfellow, “Generative adversarial networks,” ACM Digital Library, 2020.
- [18] R. G. ,. Ramin Ghorbani, “Comparing Different Resampling Methods in Predicting Students’ Performance Using Machine Learning Techniques,” Science of the Total Environment, 2020.
- [19] J. Brownlee, “Random Oversampling and Undersampling for Imbalanced Classification,” 2020. [Online]. Available: <https://machinelearningmastery.com/random-oversampling-and-undersampling-for-imbalanced-classification/>. [Accessed 28 August 2023].
- [20] F. Kamiran, “Data preprocessing techniques for classification without discrimination,” Springer Link, 2011.
- [21] IBM, “Trusted-AI/AI360,” [Online]. Available: <https://github.com/Trusted-AI/AIF360>. [Accessed 28 August 2023].
- [22] M. P. Pratik Gajane, “On Formalizing Fairness in Prediction with Machine Learning,” Cornell University, Ithaca, 2017.
- [23] Kaggle, “AI Fairness,” Google, 2023. [Online]. Available: <https://www.kaggle.com/code/alexisbcook/ai-fairness> . [Accessed 28 August 2023].
- [24] K. K. Saravanakumar, “The Impossibility Theorem of Machine Fairness -- A Causal Perspective,” Cornell University, Ithaca, 2020.
- [25] B. L. ,. M. M. Brian Hu Zhang, “Mitigating Unwanted Biases with Adversarial Learning,” Cornell University, Ithaca, NY, 2018.
- [26] . Kenna, “Using Adversarial Debiasing to remove bias from word embeddings,” arXiv , 2021.
- [27] I. Redko, E. Morvant, A. Habrard, M. Sebban and Y. Bennani, Advances in Domain Adaptation Theory, STE Press - Elsevier, 2019, p. 187.

- [28] H. S. Y. W. Shiliang Sun, “A survey of multi-source domain adaptation,” ScienceDirect, Shanghai, 2014.
- [29] G. Pearce, “Data Auditing: Building Trust in Artificial Intelligence,” [Online]. Available: <https://www.isaca.org/resources/isaca-journal/issues/2019/volume-6/data-auditing-building-trust-in-artificial-intelligence>. [Accessed 28 August 2023].
- [30] Measure Evaluation, “Data Quality Tools,” [Online]. Available: <https://www.measureevaluation.org/tools/data-quality.html>. [Accessed 28 August 2023].
- [31] W. F. ., K. M. F.-T. e. Desmond Patton, “Contextual Analysis of Social Media: The Promise and Challenge of Eliciting Context in Social Media Posts with Natural Language Processing,” ACM Digital Library, 2020.
- [32] Amazon, “Amazon Mechanical Turk,” [Online]. Available: <https://www.mturk.com/>. [Accessed 28 August 2023].
- [33] S. S. C. G. Marco Tulio Ribeiro, “Why Should I Trust You?": Explaining the Predictions of Any Classifier,” Cornell University, Ithaca, 2016.
- [34] D. B. ., S. F. S. V. ., J. V. Andrew D. Selbst, “Fairness and Abstraction in Sociotechnical Systems,” ACM Conference on Fairness, Accountability, and Transparency , Los Angeles, 2018.
- [35] Vanilla Forums, “Open Source Community Forum Software,” 2023.
- [36] Vanilla Forums, “the Success Community,” [Online]. Available: <https://success.vanillaforums.com/kb>. [Accessed 2023 August 28].
- [37] alexisbcook, “Kaggle/learntools,” [Online]. Available: <https://github.com/Kaggle/learntools/tree/master/notebooks/ethics/pdfs>. [Accessed 28 August 2023].
- [38] Framer, “What is Framer?,” [Online]. Available: <https://www.framer.com/learn/what-is-framer/>. [Accessed 28 August 2023].
- [39] Framer, “Site Pricing,” [Online]. Available: <https://www.framer.com/pricing/>. [Accessed 28 August 2023].
- [40] AWS, “Amazon Route 53,” [Online]. Available: <https://aws.amazon.com/route53/>. [Accessed 28 August 2023].
- [41] AWS, “What is an Application Load Balancer?,” [Online]. Available: <https://docs.aws.amazon.com/elasticloadbalancing/latest/application/introduction.html>. [Accessed 28 August 2023].

- [42] AWS, “Provide network connectivity for your Auto Scaling instances using Amazon VPC,” [Online]. Available: <https://docs.aws.amazon.com/autoscaling/ec2/userguide/asg-in-vpc.html>. [Accessed 28 August 2023].
- [43] Aviatrix, “What is Azure Network Virtual Appliance (NVA)?,” [Online]. Available: <https://aviatrix.com/learn-center/cloud-security/azure-network-virtual-appliance/>. [Accessed 28 August 2023].
- [44] AWS, “How to integrate third-party firewall appliances into an AWS environment,” [Online]. Available: <https://aws.amazon.com/blogs/networking-and-content-delivery/how-to-integrate-third-party-firewall-appliances-into-an-aws-environment/>. [Accessed 28 August 2023].
- [45] AWS, “AWS Identity and Access Management,” [Online]. Available: <https://aws.amazon.com/iam/>. [Accessed 28 August 2023].
- [46] AWS, “Amazon RDS for PostgreSQL,” [Online]. Available: <https://aws.amazon.com/rds/postgresql/>. [Accessed 28 August 2023].
- [47] AWS, “Elastic Load Balancing,” [Online]. Available: <https://aws.amazon.com/elasticloadbalancing/>. [Accessed 28 August 2023].
- [48] AWS Transit Gateway, [Online]. Available: <https://aws.amazon.com/transit-gateway/>. [Accessed 28 August 2023].
- [49] AWS, “AWS WAF,” [Online]. Available: <https://aws.amazon.com/waf/>. [Accessed 28 August 2023].
- [50] OWASP, “Attacks,” [Online]. Available: <https://owasp.org/www-community/attacks/>. [Accessed 28 August 2023].
- [51] AWS, “What is Amazon EC2 Auto Scaling?,” [Online]. Available: <https://docs.aws.amazon.com/autoscaling/ec2/userguide/what-is-amazon-ec2-auto-scaling.html>. [Accessed 28 August 2023].
- [52] AWS, “Amazon EC2 C5 Instances,” [Online]. Available: <https://aws.amazon.com/ec2/instance-types/c5/>. [Accessed 28 August 2023].
- [53] AWS, “AWS Lake Formation,” [Online]. Available: <https://aws.amazon.com/lake-formation/>. [Accessed 28 August 2023].
- [54] AWS, “Amazon S3,” [Online]. Available: <https://aws.amazon.com/s3/>. [Accessed 28 August 2023].

- [55] Hopsworks, “Guide to File Formats for Machine Learning,” [Online]. Available: <https://www.hopsworks.ai/post/guide-to-file-formats-for-machine-learning>. [Accessed 28 August 2023].
- [56] AWS, “Machine Learning Frameworks and Languages,” [Online]. Available: <https://docs.aws.amazon.com/sagemaker/latest/dg/frameworks.html>. [Accessed 28 August 2023].
- [57] AWS, “Data cataloging,” [Online]. Available: <https://docs.aws.amazon.com/whitepapers/latest/best-practices-building-data-lake-for-games/data-cataloging.html>. [Accessed 28 August 2023].
- [58] AWS, “Effective data lakes using AWS Lake Formation, Part 2: Securing data lakes with row-level access control,” [Online]. Available: <https://aws.amazon.com/blogs/big-data/part-2-effective-data-lakes-using-aws-lake-formation-secure-data-lakes-with-row-level-access-control/#:~:text=Lake%20Formation%20row%2Dlevel%20permissions,when%2C%20and%20through%20which%20services..> [Accessed 28 August 2023].
- [59] AWS, “What Is AWS CloudTrail?,” [Online]. Available: <https://docs.aws.amazon.com/awscloudtrail/latest/userguide/cloudtrail-user-guide.html>. [Accessed 28 August 2023].
- [60] AWS, “Amazon API Gateway,” [Online]. Available: <https://aws.amazon.com/api-gateway/>.
- [61] AWS, “AWS Lambda,” [Online]. Available: <https://aws.amazon.com/lambda/>. [Accessed 28 August 2023].
- [62] AWS, “AWS Lambda Features,” [Online]. Available: <https://aws.amazon.com/lambda/features/>. [Accessed 28 August 2023].
- [63] AWS, “AWS Step Functions,” [Online]. Available: <https://aws.amazon.com/step-functions/>. [Accessed 28 August 2023].
- [64] AWS, “AWS Glue,” [Online]. Available: <https://aws.amazon.com/glue/>. [Accessed 28 August 2023].
- [65] AWS, “What is Amazon Athena?,” [Online]. Available: <https://docs.aws.amazon.com/athena/latest/ug/what-is.html>. [Accessed 28 August 2023].
- [66] AWS, “Control and audit data exploration activities with Amazon SageMaker Studio and AWS Lake Formation,” [Online]. Available: <https://aws.amazon.com/blogs/machine-learning/controlling-and-auditing-data-exploration-activities-with-amazon-sagemaker-studio-and-aws-lake-formation/>.
- [67] AWS, “What is Amazon SageMaker?,” [Online]. Available: <https://docs.aws.amazon.com/sagemaker/latest/dg/whatis.html#how-it-works>. [Accessed 28 August 2023].

- [68] AWS, “AWS Pricing Calculator,” [Online]. Available: <https://calculator.aws/#/>. [Accessed 28 August 2023].
- [69] Glassdoor, “Senior Software Engineer Salaries,” [Online]. Available: https://www.glassdoor.com/Salaries/south-africa-senior-software-engineer-salary-SRCH_IL.0,12_IN211_KO13,37.htm. [Accessed 28 August 2023].
- [70] Glassdooe, “How much does a Senior Data Scientist make in Johannesburg, South Africa?,” [Online]. Available: https://www.glassdoor.com/Salaries/johannesburg-senior-data-scientist-salary-SRCH_IL.0,12_IM1023_KO13,34.htm. [Accessed 28 August 2023].
- [71] payscale, “Average Social Scientist Salary in South Africa,” [Online]. Available: https://www.payscale.com/research/ZA/Job=Social_Scientist/Salary. [Accessed 28 August 2023].
- [72] Glassdoor, “How much does a Marketing Manager make in Johannesburg,” [Online]. Available: https://www.glassdoor.com/Salaries/johannesburg-marketing-manager-salary-SRCH_IL.0,12_IC2638926_KO13,30.htm. [Accessed 28 August 2023].
- [73] Economic Research Institute, “Machine Learning Research Scientist Salary,” [Online]. Available: <https://www.erieri.com/salary/job/machine-learning-research-scientist/south-africa>. [Accessed 28 August 2023].
- [74] J. Barr, “New – Customer Carbon Footprint Tool,” AWS, [Online]. Available: <https://aws.amazon.com/blogs/aws/new-customer-carbon-footprint-tool/>. [Accessed 28 August 2023].
- [75] The Climate Pledge, “Be the planet's turning point,” [Online]. Available: <https://www.theclimatepledge.com/#main-navigation>. [Accessed 28 August 2023].

Appendix A: Short non-technical report

Artificial Intelligence (AI) is swiftly finding its way into pivotal decision-making roles across society. Its remarkable capabilities are fostering a transformative landscape, brimming with potential avenues for progress. Nevertheless, as these advancements hold immense promise, it's crucial to understand the complex processes through which AI systems make decisions and to rigorously evaluate the results they produce. In light of this landscape, the aim of this design initiative revolves around conceptualizing and establishing an innovative online platform geared towards addressing the negative impacts of innate biases within AI models.

Bias, in machine learning, refers to the presence of systematic errors or unfairness in the predictions and decisions made by AI models, often leading to discriminatory and inaccurate outcomes. Recognizing the inherent challenge of dealing with bias at every step of the AI process, this design emphasizes the importance of grasping the context in identifying possible sources of unfairness. With this awareness in mind, a meticulous design solution was crafted, culminating in an intuitive user interface that systematically guides users through the process of uploading datasets, models, and contextual answers. In essence, this interface has been carefully designed to ensure its accessibility, even for those without advanced proficiency in machine learning terminology, thus enabling widespread engagement with the platform's functionalities.

Users are reassured about uploading this sensitive and personal data due to the robust security and privacy measures designed within the web application infrastructure. As each upload takes place, advanced security applications provided by Amazon Web Services (AWS) - the largest and a reputable cloud infrastructure provider - analyse and protect the information. Throughout the storage and maintenance of the data, vigorous security and privacy measures are implemented. Access is granted to only specific roles within the company, and all data is tracked, maintaining records that can be monitored and audited.

Likewise, the use of AWS allows the platform to scale based on user activity and storage requirements. This, in turn, enables ad hoc pricing mechanisms for each specific user. As a result, the platform offers a versatile pay-as-you-go model, making it accessible to small, medium, and large companies. Therefore, since biased model can affect industries across the board, irrespective of a company's size, the bias mitigation application is well equipped to adeptly tackle and alleviate these adverse effects stemming from biased AI models across all scenarios.

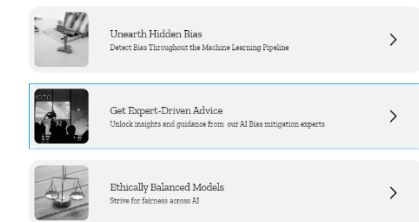
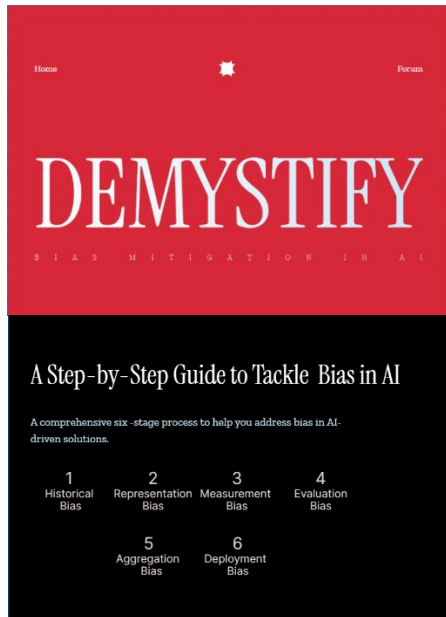
Furthermore, upon uploading datasets and models, as well as filling in the contextual framework within the online platform, users gain a comprehensive understanding of bias instances throughout the AI journey. Subsequently, a meticulously crafted report is generated, overseen by a team of engineers, social scientists, and expert machine learning researchers who collaborate to ensure precision. This infusion of human expertise not only boosts user confidence but also fosters a design prioritizing fairness, transparency, and accountability, catering effectively to diverse user groups.

The curated report delivers an insightful context overview, encompassing the user-provided details. Additionally, it incorporates personalized visualizations employing advanced techniques to highlight critical aspects of the data and model. These insights are further clarified to the user through fairness metrics, explaining the statistical perpetuation of bias within the data and model. Armed with these metrics, the web application provides expert guidance via bias mitigation techniques to rectify the adverse effects arising from biased AI models.

In addition to the visualizations and actionable insights, environmental sustainability is a core value embedded in the curated report. By utilizing AWS's carbon footprint tracking capabilities, the user is provided with key insights into their carbon emissions. These statistics demonstrate how much energy has been saved through the use of renewable energy and cutting-edge computing resources. Therefore, by designing the bias mitigation platform with an eco-conscious approach, users are provided with a sustainable solution that enhances its value in addressing machine learning bias throughout society.

Nevertheless, in cases where users remain unconvinced about bias mitigation, an option for crowdsourcing emerges through a dedicated forum hosted by the web application. This forum encourages users to offer feedback and share their model deployment experiences, fostering an environment for collective learning and continuous improvement. Conclusively, this all-encompassing approach empowers users with profound insights into the nuances of bias mitigation, forging a collaborative space that fosters ongoing dialogue, encapsulating the very essence of a user-centric and ever-evolving solution aimed at mitigating the negative impacts of biased AI across society.

Appendix B: Additional Tables and Figures



Frequently Asked Questions

How is the pricing structured for your services?

Our pricing is flexible, customized to your needs, it factors in variables like model size, complexity, and the required AWS data infrastructure. Contact us to discuss your project, and we'll provide you with a personalized pricing estimate.

What kind of AI models do you support?

Our model-agnostic process is tailored to suit AI models across various domains including healthcare, finance, and technology.

Do I need Machine Learning expertise?

No, our platform is designed to guide you through the entire process, making it accessible for all users.

How do you ensure user data privacy?

We take privacy and data security seriously. All uploaded user data is encrypted and stored securely. We follow industry standards and comply with data protection regulations. Your data is only used for bias assessment and mitigation, and we never share or sell personal information.

Ready to begin the journey to AI fairness? Let's start mitigating bias today!

[Get started](#) [Learn more](#)

© Bias Mitigator 2023

(a)



What is Historical Bias?

Historical bias, a key factor in bias within machine learning, arises when data originates from a flawed past environment

Kindly provide context-specific answers to the following questions:

1. Could you please provide details about the source and origin of the historical data, along with insights into the context and methodology of its collection?

Submit

2. Has the personal information been obtained in a consensual fair manner in accordance with the POPIA act?

Submit

3. Does the dataset encompass any specific protected demographic attributes, such as gender, ethnicity, or age, which have historically been linked to bias and discrimination? Please identify any attributes that might influence bias.

Submit

4. Could you please provide more details about the target variable or prediction you are aiming to achieve with this dataset? Please elaborate on the specific outcome you are trying to predict.

Submit

5. Which specific metric or threshold would you prefer to be used for analyzing the results of the assessment? Please indicate the measurement criteria that align with your evaluation preferences.

Submit

Please upload your dataset:

Upload

Please upload your model :

Upload

[Submit](#)

© Bias Mitigator 2023

(b)

Figure 2: UI Landing Page: <https://mitigating-bias-in-ai.framer.ai/>, Historical Bias UI: <https://mitigating-bias-in-ai.framer.ai/HistoricalBias>

Home
Form

REPRESENTATION BIAS

BIAS MITIGATION IN AI

What is Representation Bias?

Representation Bias occurs when the datasets used to train a model do not accurately reflect the people the model is meant to help.

Kindly provide context-specific answers to the following questions:

- Do you believe the individuals in your training dataset truly representative of the diverse population that your model will serve?

Details of training data
Submit

- Can you describe the data collection process and how samples were obtained? Are there any limitations in data collection that might lead to representation bias?

Limitations of data collection
Submit

- Does your training data accurately reflect the current characteristics and trends of the population you intend to apply the model to?

Current characteristics
Submit

- Do you know if distribution of your training data aligned with the distribution of the real-world data where your model will be deployed?

Details about distribution
Submit

- Have you considered using adversarial debiasing techniques to enhance fairness in your model?

Yes or No
Submit

Submit



© Bias Mitigator 2023

Home
Form

EVALUATION BIAS

BIAS MITIGATION IN AI

What is Evaluation Bias?

Evaluation bias occurs when evaluating a model and the benchmark data does not represent the population that the model will serve.

Kindly provide context-specific answers to the following questions:

- Could you describe the origin and composition of your benchmark dataset? Specifically, have any historical, representational, or measurement biases been identified or suspected in the data collection process?

Details about Benchmark Dataset
Submit

- Are there any known limitations or biases associated with the benchmark dataset that could potentially impact the evaluation of machine learning models?

Details about Data Generation...
Submit

- How do you plan to compare different models using the benchmark dataset? Are there any specific performance metrics you are aiming to evaluate?

Details about how the Benchmark data is compared
Submit

- Have you identified any challenges or concerns related to improper comparison methods that could potentially introduce evaluation bias during model assessment?

Details about Challenges regarding benchmark datasets
Submit

- Are there any existing or published flaws in the benchmark dataset that you are aware of? Additionally, how open are you to the proposed design's approach of subjecting the benchmark data to mitigation techniques and fairness evaluation to ensure its validity?

Details about existing known flaws
Submit

Dataset Upload
Upload

Submit



© Bias Mitigator 2023

Figure 3: Representation Bias UI: <https://mitigating-bias-in-ai.framer.ai/RepresentationBias>, Evaluation Bias UI: <https://mitigating-bias-in-ai.framer.ai/EvaluationBias>




(a)



(b)

Figure 4: Aggregation Bias UI: <https://mitigating-bias-in-ai.framer.ai/AggregationBias> , Measurement Bias UI: <https://mitigating-bias-in-ai.framer.ai/MeasurementBias>

[Home](#)



[Forum](#)

DEPLOYMENT BIAS BIAS

B I A S M I T I G A T I O N I N A I

What is Deployment Bias?

Deployment bias arises when the model's intended purpose diverges from its actual usage scenario.

Kindly provide context-specific answers to the following questions:

1. What is the primary objective or problem that your AI model is designed to address?

Submit

2. Could you describe the typical scenarios or situations in which your AI model will be deployed or used?

Submit

3. Are there any specific factors or conditions in the real-world usage of the model that might differ from the ideal scenario you initially had in mind?

Submit

4. How do you anticipate potential differences between the intended use case of the model and its actual deployment might impact the model's performance or effectiveness?

Submit

We highly encourage users to actively engage in providing valuable feedback and fostering meaningful discussions about bias mitigation. Join the conversation and share your insights by clicking [here](#) to access the forum hosted by Vanilla Forums. Your input is essential in enhancing the effectiveness and impact of the bias mitigation techniques discussed.

Submit




Figure 5: Deployment Bias UI: <https://mitigating-bias-in-ai.framer.ai/DeploymentBias>

26

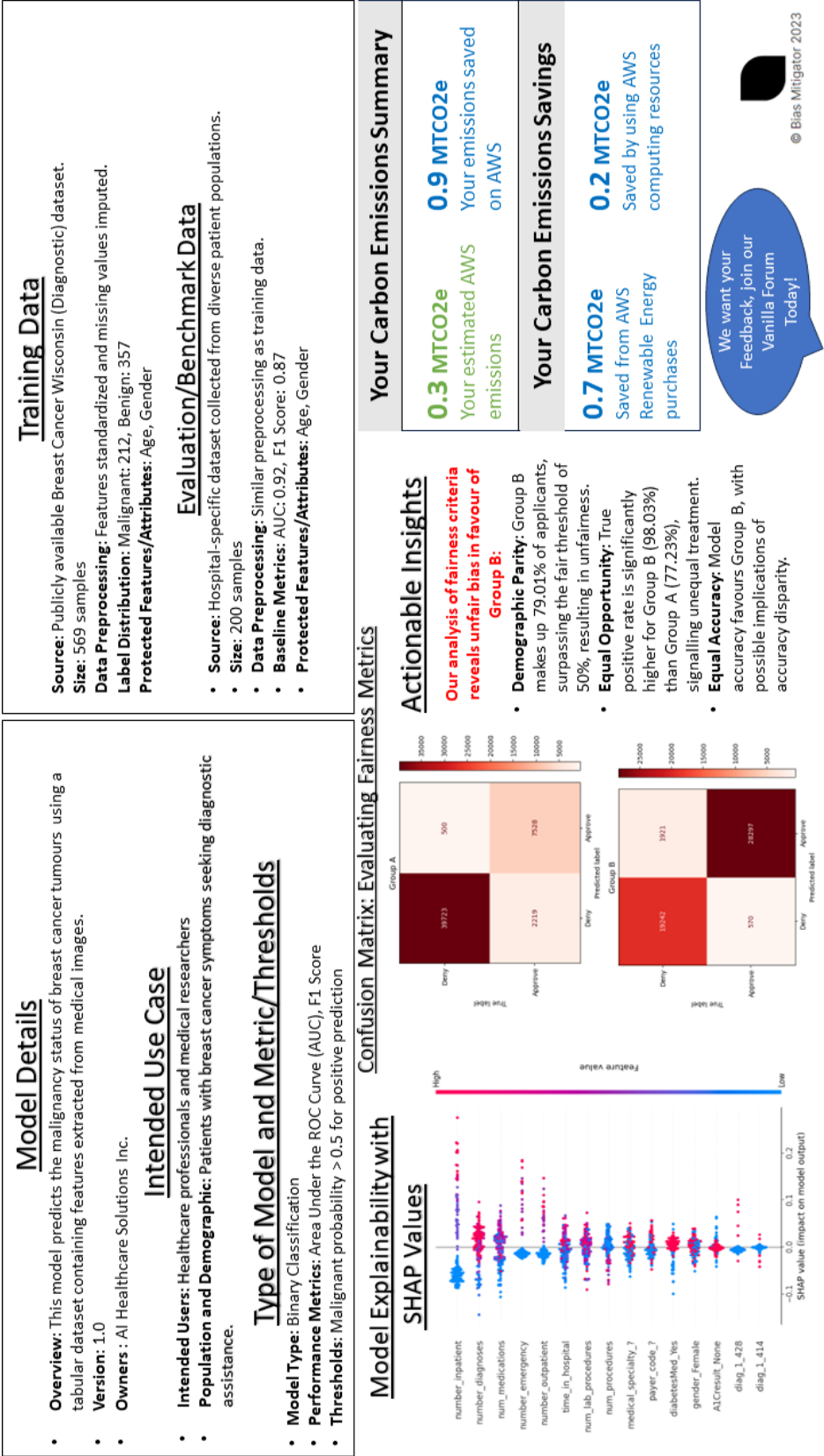


Figure 6: Curated Report providing context and insights for bias mitigation.

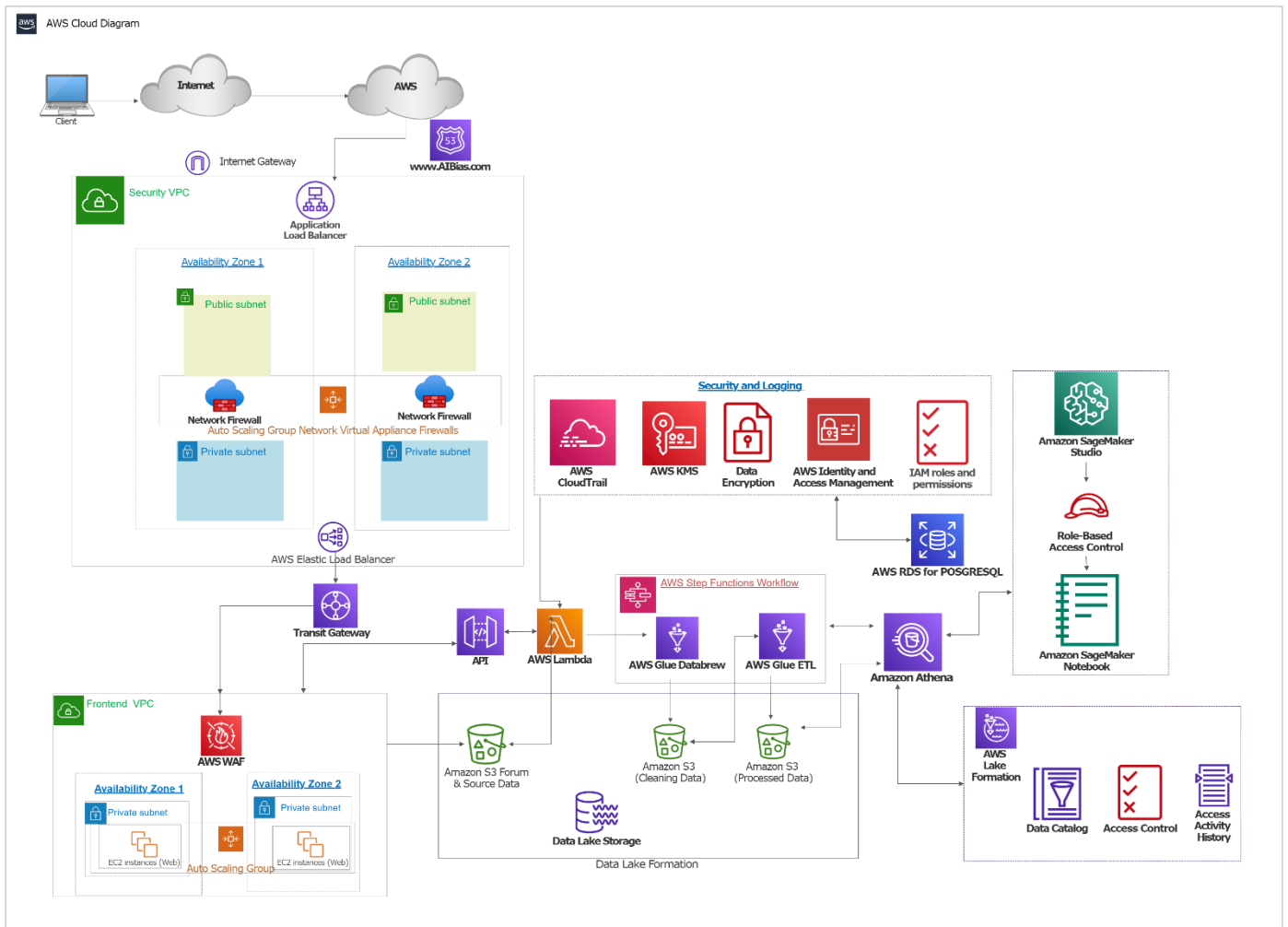


Figure 7: Cloud Infrastructure Design

[illegible]