RV UNIVERSITY, BENGALURU-59

SCHOOL OF COMPUTER SCIENCE AND ENGINEERING

A Mini Project Report On

Extraction of Dark data from Electronic Health Records using Ensemble models

Submitted in partial Fulfillment for the award of degree of

B. Tech (Honors)

In

School of Computer Science and Engineering

Submitted By

Jeevan E G	1RVU22CSE069
Kalp Jain	1RVU22CSE076
Pranav P Kulkarni	1RVU22CSE119

Under the Guidance of

Prof.Sonam V Maju

Assistant Professor

School of CSE

RV University, Bengaluru-560059

2024-2025

RV UNIVERSITY, BENGALURU-59

SCHOOL OF COMPUTER SCIENCE AND ENGINEERING

CERTIFICATE

Certified that the mini project work titled **“Extraction of Dark data from Electronic Health Records using Ensemble models** “is carried out by **Jeevan EG-1RVU22CSE069, Kalp Jain-1RVU22CSE076, Pranav P Kulkarni-1RVU22CSE119** who are bonafide students of RV University, Bengaluru, in partial fulfilment of **Bachelor of Technology (Hons) in School of Computer Science and Engineering** of the RV University, Bengaluru during the year 2025-2026. It is certified that all corrections/suggestions indicated for the Internal Assessment have been incorporated in the mini project report deposited in the departmental library. The Mini Project report has been approved as it satisfies the academic requirements in respect of mini project work prescribed by the institution for the said degree.

Signature of Guide **Guide Name**	**Signature of Program Director** **Dr. Sudhakar K N**	**Signature of Dean** **Dr.Shobha G**

External Viva:

Name of Examiners	**Signature with Date**
1	
2	

RV UNIVERSITY, BENGALURU-59

SCHOOL OF COMPUTER SCIENCE AND ENGINEERING

DECLARATION

We, Jeevan E G, Kalp Jain, and Pranav P Kulkarni students of sixth semester B. Tech (Hons), SoCSE, RV University, Bengaluru, hereby declare that the mini project titled **'Extraction of Dark data from Electronic Health Records using Ensemble models'** has been carried out by us and submitted in partial fulfilment of **Bachelor of Technology (Hons)** in **School of Computer Science and Engineering during** the year 2025-26.

Further we declare that the content of the report has not been submitted previously by anybody or to any other university.

We also declare that any Intellectual Property Rights generated out of this project carried out at RV University will be the property of RV University, Bengaluru and we will be one of the authors of the same.

Place: Bengaluru

Date:

Name **Signature**

1. Jeevan E G-1RVU22CSE069

2. Kalp Jain-1RVU22CSE076

3. Pranav P Kulkarni-1RVU22CSE119

ACKNOWLEDGEMENT

It is a great pleasure for us to acknowledge the assistance and support of many individuals who have been responsible for the successful completion of this project work.

First, we take this opportunity to express our sincere gratitude to School of Computer Science and Engineering, RV University for providing us with a great opportunity to pursue our bachelor’s degree in this institution.

A special thanks to our Program Director **Dr. Sudhakar, Dean Dr. Shobha G** for their continuous support and providing the necessary facilities with guidance to carry out mini project work.

We would like to thank our guide **Prof.Sonam V Maju**, **Assistant Professor**, **School of Computer Science and Engineering, RV University,** for sparing his/her valuable time to extend help in every step of our project work, which paved the way for smooth progress and fruitful culmination of the project.

We are also grateful to our family and friends who provided us with every requirement throughout the course.

We would like to thank one and all who directly or indirectly helped us in the Project work.

Signature of Student

USN:

Name:

Abstract

Purpose:

The objective of this mini project is to extract valuable hidden insights, commonly referred to as dark data, from synthetic Electronic Health Records (EHR) related to blood cancer patients. The focus is on integrating both structured and unstructured medical data to uncover patterns that are often overlooked in traditional clinical analysis.

Methodology:

Synthetic patient data was initially generated in JSON format, containing a wide range of demographic, diagnostic, treatment, and laboratory attributes. After preprocessing and converting the data into CSV format, it was separated into structured (patient attributes) and unstructured (doctor notes and symptom descriptions) components. Structured data was processed using ensemble classification models, targeting the treat_outcome attribute as the prediction label. Simultaneously, unstructured data was analysed using advanced Natural Language Processing (NLP) techniques, employing the ScispaCy’s en_ner_bc5cdr_md model. model for extracting key medical entities and contextual patterns. The outcomes from both streams were combined to generate a comprehensive JSON file representing the final extracted dark data.

Results:

The project successfully identified hidden patterns and connections within the dataset, enhancing the interpretability of patient conditions and treatment responses. The integrated results provided a more holistic view of patient records, revealing associations between therapies, genetic markers, and symptom patterns that were not directly visible through conventional analysis.

Conclusion:

This work demonstrates the potential of combining machine learning and NLP techniques to extract critical, underutilized information from healthcare data, contributing to more informed clinical decision-making and research advancements.

Table of Contents

List of Tables

List of Figures

Chapter-1: Introduction

In the last recent years, the health-care industry has been implementing Electronic Health Records (EHR) to capture patient related information electronically. These EHRs can consist of components such as patient demographics, laboratory reports, diagnosis, any treatments, and other clinical notes. Most of this data is structured with a consistent data format and has been actively mined, analysed, and reported. However, data that is classified as "dark data" exists, which is data has been captured, but not used or analyzed. Dark data may consist of incomplete data records, missing and hidden data patterns, and especially data captured in the unstructured text format such as doctors' prescriptions. Mining and analyzing this dark data could offer major contributions to improved quality of health-care, improve decision-making in hospital wards and primary patient health care practices. [1]

This mini-project is going to create synthetic EHR data, with structured attributes and a doctors' prescription column with unstructured data. To mine for missing and hidden alternative patterns in this dark data, we have made use of ensemble machine learning models such as Random Forest, Gradient Boosting, and XGBoost. Moreover, we applied Natural Language Processing (NLP) techniques to the prescription notes to extract medical terms, symptoms, medications, and any other relevant information required to be valid day-to-day appropriate medical decisions when working for patients in clinical settings. By analysis structured data, using machine learning and unlocking dark data with the use of the unstructured text through data mining and NLP techniques, we will be demonstrating how we can identify dark data and transform its meanings and use within environments of this nature. [2]

1.1 State of Art Development

The ability to extract hidden or underused information (the euphemistic term is "dark data") from Electronic Health Records (EHRs) has evolved as an important topic in health analytics. Recent developments have also included ensemble learning methods and Natural Language Processing (NLP) methods to retrieve these hidden insights. Ensemble methods, which combine multiple learning methods, have shown improved accuracy in tasks like de-identifying clinical narratives and retrieving a variety of information pieces from clinical notes. For example, combining cTAKES and MetaMap in ensemble methods has improved the identification of clinical concepts in a variety of clinical conditions.

At the same time, categories of NLP have been formed to handle collections of free text and unstructured textual data in EHRs, such as physician notes and prescriptions. The utilization of Named Entity Recognition (NER) was implemented to obtain clinical specifics, and therefore assist in predicting disease and stratifying patients. The models Med-BERT and Hi-BEHRT, have been created for medical domain data using transformer architectures, which is another step forward for predicting clinical events and resolving the course of patients. Overall, these facilities contribute to a greater use of EHRs and the ability of healthcare providers to furnish clinical decisions that are the product of larger patient data sets. [3]

1.2 Motivation

In today's health care landscape Electronic Health Records (EHRs) have become a necessary hassle for managing patients' data, with structured data like demographics, lab values, etc. being reviewed oftentimes. However, fully a third of the data, or unstructured or unused data (aka dark data) is buried in physician prescriptions, clinical notes, and treatment history and contains valuable insight that may offer better patient decision making and care. Our motivation for this project is to tap into that data and retrieve its hidden insight. By generating synthetic EHR data, we can pair ensemble machine learning models along with Natural Language Processing (NLP) methods to potentially identify patterns from both organic structured and unstructured data. This project allows us to extend our insights of clinical data processing through the use of synthetic patient data and to exemplify one-way advanced AI methods and approaches may be used to improve healthcare analytics in a privately protected, research-driven environment. [4]

1.3 Problem Statement

While Electronic Health Record systems are widely utilized across the healthcare industry, much data remains unstructured and untapped, particularly in sections like doctor's prescriptions and narrative reports - unstructured data which may even contain valuable clinical information which might provide better predictions of healthcare outcomes, better risk profiles, and more targeted patient interventions under the right extraction and analysis processes.

The challenge is multi-faceted: to process the unstructured data associated with health records together with structured data; to identify relevant hidden data; and to generate usable insights. Single-model approaches typically struggle to comprehend complexity in diverse data.

Therefore, there is a need for a system that combines multiple models and techniques to accurately extract and interpret this hidden, or "dark," data within EHRs. [5]

1.4 Objectives

This study aims to create a synthetic dataset mimicking real Electronic Health Records (EHR) with a degree of structured as well as unstructured data. The structured component will address factors like the patient's details, laboratory test reports, and their history of diseases. This data will be used with ensemble machine learning models in an attempt to spot patterns not immediately perceivable. Simultaneously, the unstructured elements—like handwritten physician prescriptions or notes—will be processed using Natural Language Processing (NLP) techniques to yield useful information. Furthermore, this study will explore whether ensemble models perform better than the traditional single models when working on EHR data. Finally, the goal is to show how advanced AI methods can help uncover concealed insights in healthcare data and support making smarter, more accurate clinical decisions.. [6]

1.5 Methodology

The methodology followed in this project is designed to systematically extract valuable hidden or "dark" data from synthetic Electronic Health Records (EHR) using ensemble machine learning models and Natural Language Processing (NLP) techniques. The following steps summarize the process undertaken:

A. Pre-Processing:

At first, the EHR data set was created in JSON format, which included both structured information (e.g. demographics, diagnosis, lab values) and unstructured information (like doctor's orders). In order to make the management and analysis easier, the JSON data set was converted to CSV format.

Once in CSV format, the structured data was cleaned and pre-processed. This included:

- Handling missing values by replacing null or empty fields appropriately.
- Converting categorical data into numerical representations using label encoding and one-hot encoding wherever necessary.
- Normalizing numerical attributes for consistency and better model performance.

- Separating the unstructured text (doctor’s prescription column) from the structured data for dedicated NLP processing.

B. Text Extraction:

The unstructured data from doctors' prescriptions was processed using Natural Language Processing (NLP) techniques in the following steps: unnecessary characters, punctuation, and stopwords were removed; tokenization and lowercasing were applied to standardize the text data; NLP-based keyword extraction techniques were employed to identify key medical terms and clinical indicators; and finally, the cleaned text was transformed into numerical data using term frequency-inverse document frequency (TF-IDF) vectorization to make it suitable for machine learning models. [7]

C. Application:

After pre-processing and text extraction, both structured and unstructured data were analyzed using ensemble machine learning models. The analysis involved employing an ensemble learning approach that combined random forest, boosting methods, and voting classifiers to improve accuracy and handle complex datasets effectively. These models were trained on the processed data to uncover hidden patterns, insights, or data artifacts. The evaluation of the models was conducted using various metrics such as accuracy, precision, recall, F1 score, significance measures, ranks, and identification tables or charts. Finally, the results and insights were visualized, illustrating that ensemble models significantly outperformed single-model approaches in extracting valuable hidden data from synthetic Electronic Health Records (EHR). This systematic methodology allowed for effective handling of both structured and unstructured healthcare data, successfully demonstrating the potential of ensemble learning and NLP in uncovering hidden clinical insights. [8]

1.6 Innovation

One of the broader innovations of this project is that it has generated a full synthetic, tailored Electronic Health Record (EHR) data set that comprised 500 patient records and 51 distinctive attributes. This project diverges from other conventional research employing publicly and/or existing available clinical data, as it permitted researchers to bypass the challenge of small medical databases, and develop synthetic and full records (which were designed) with a view to extracting and triangulating dark data in blood cancer.

The dataset included a holistic set of patient data, namely demographic data (age, sex, race, occupation, etc.), medical data (family cancer history, previous diagnoses, interventions, and birth information), as well as data on laboratory investigations (e.g., results from blood tests such as WBC, RBC, hemoglobin levels, etc.). It also incorporated genetic mutation markers (e.g., Philadelphia chromosome, FLT3 mutation, TP53 mutation, etc.), unstructured data presented in the form of doctor orders and comments (essential for developing databases), and treatment plans, medications, outcomes of patients, and social determinants such as smoking, drinking, and exercising habits.

One of the most distinct components of this work is including free-text 'doctor_notes' and 'side_effects_text' fields, because it provides the ability to take advantage of Natural Language Processing (NLP) techniques to gain knowledge that isn't apparent from unstructured clinical text. It is a vital inclusion because most artificial datasets do not have the option of any purely structured fields either.

The data set was also initially in JSON format, which appeared to provide flexibility with the structuring of data, and was able to hold hierarchical relations. But subsequently it was converted into csv to take advantage of machine learning libraries, and for convenience of handling the data set during model training and testing. This approach not only demonstrates the feasibility of synthesizing realistic healthcare data but also provides a controlled environment to evaluate ensemble machine learning models and NLP techniques for dark data extraction — an under-investigated combination in existing work.

1.7 Organization of the report

This report is structured to provide a detailed overview of the mini project titled “Extraction of Dark Data from EHR”. It systematically describes the motivation, methodology, implementation, and outcomes of the project, offering insights into both structured and unstructured healthcare data analysis.

- **Chapter 1: Introduction**
  This chapter introduces the background, motivation, objectives, and scope of the mini project, emphasizing the need for extracting hidden patterns from healthcare records.

- **Chapter 2: Literature Survey**
  It presents a review of existing techniques and tools used in the extraction of dark data, along with research works on EHR analysis, ensemble models, and NLP applications.
- **Chapter 3: System Analysis**
  This section outlines the problem definition, feasibility study, and system requirements for implementing the proposed solution effectively.
- **Chapter 4: System Design**
  It describes the architecture and design of the system, including data flow, components, and methodologies chosen for both structured and unstructured data processing.
- **Chapter 5: Implementation**
  This chapter details the implementation process of data preprocessing, classification models, ScispaCy's en_ner_bc5cdr_md model and integration of results to extract dark data.
- **Chapter 6: Results and Discussion**
  It presents the key findings obtained after applying the proposed methodology, along with a discussion on their relevance, accuracy, and insights generated.
- **Chapter 7: Conclusion and Future Work**
  This chapter concludes the report by summarizing the outcomes of the project and suggesting possible directions for further enhancement and real-world applications.
- **Appendices**
  It includes additional materials such as datasets, sample outputs, code snippets, and documentation that support the content of the report.

Chapter-2: Overview of Project

The diagram outlines a novel data architecture method for extracting dark data from blood cancer electronic health records. The pipeline accounts for the complex nature of medical information with a dual analysis, structured data goes through ensemble classification by focusing on each therapy indicator; while unstructured clinical narratives use ScispaCy's natural language model, en_ner_bc5cdr_md. The pipeline begins with the pre-structured JSON data made synthetically with 51 clinical data attributes. The pipeline first pre-processes, separates, and analyzes the two distinct but complementary data types before concatenating the data to find the previously unknown associations and clinically relevant patterns missed by traditional single-method analyses. [9]

2.1 Diagram of your project and Introduction

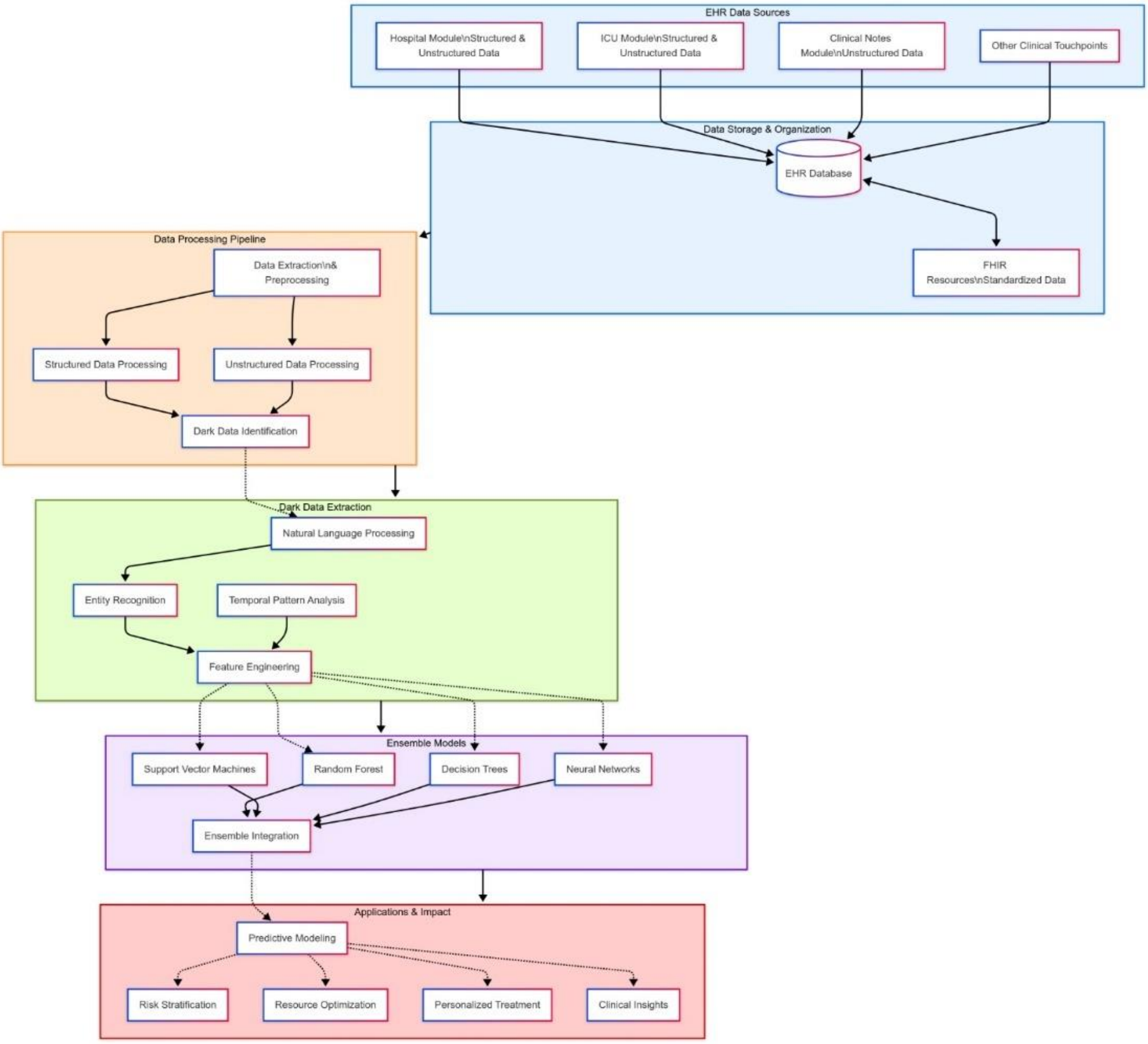

EHR Data Sources
Hospital Module\nStructured & Unstructured Data
ICU Module\nStructured & Unstructured Data
Clinical Notes Module\nUnstructured Data
Other Clinical Touchpoints
Data Storage & Organization
EHR Database
FHIR Resources\nStandardized Data
Data Processing Pipeline
Data Extraction\n& Preprocessing
Structured Data Processing
Unstructured Data Processing
Dark Data Identification
Dark Data Extraction
Natural Language Processing
Entity Recognition
Temporal Pattern Analysis
Feature Engineering
Ensemble Models
Support Vector Machines
Random Forest
Decision Trees
Neural Networks
Ensemble Integration
Applications & Impact
Predictive Modeling
Risk Stratification
Resource Optimization
Personalized Treatment
Clinical Insights

Figure 1. System Overview of Dark Data Extraction from Electronic Health Records Using Ensemble Models

Chapter-3 Software Requirements

This chapter has thoroughly detailed the functional and non-functional requirements as well as the relevant hardware and software requirements that will be important for the proper completion of the mini-project. The requirements specified here were based on the scope of the system, the complexity of the data involved, and the computational demands of executing ensemble methods and NLP methods on synthetic Electronic Health Record (EHR) data. [10]

3.1 Functional Requirements

The scope of the project specifies the system's primary use cases. This includes generation and processing of synthetic electronic health records (EHR) datasets in JSON format, which later undergoes modeling in CSV format. The system will also include preprocessing of both structured and unstructured data like side effect text, physician notes, and lab work. We'll utilize Natural Language Processing (NLP) methods for extracting valuable information out of the unstructured text sections of the data, such as physician's comments or prescriptions. For discovering concealed patterns and achieving precise predictions, ensemble machine learning models like Random Forest, Gradient Boosting, and Voting Classifiers will be used. Besides that, the entire system will be capable of providing visual and explanatory disseminations of concealed insights and outcomes from model simulations. Moreover, EHR data when modified will be stored and retrieved without queries being made at the time and datasets validated for coherence and consistency throughout the processing stages**.**[11]

3.2 Non-Functional Requirements

Like the functional requirements, these objectives also aim at obtaining system goals. These objectives involve important characteristics such as scalability, which is the ability of the system to grow by accommodating additional datasets such as patient records or attributes. Another characteristic that is crucial is precision, which in our context refers to accuracy with respect to providing output for pattern recognition and drawing insights using ensemble models and NLP techniques.

Another area is failure, which is the reliable expectation that a platform will deal with large volumes of synthetic medical data with minimal failure. There is high regard for constrain maintainability which refers to the code and datasets being easily alterable to allow subsequent changes, for instance setting in additional attributes or new ML algorithms. Easy to use, well documented, well visualized if not add visually interpretable, especially and centrally for interpretation, also put in the spotlight which helps to conceived as usability.

The last of the performance objectives rests on the system being able to provide a reasonable response time that is proportional to the time taken to process a set of data based on the complexity and size of the EHR dataset. [12]

3.3 Hardware Requirements

The hardware requirements formulated for the project were:

- Processor: Intel Core i5 or higher (or equivalent AMD Ryzen 5 series)
- RAM: Minimum 8 GB (16 GB recommended for faster processing)
- Storage: At least 5 GB of free space for datasets, libraries, and outputs
- Operating System: Windows 10/11 or Linux (WSL for Linux-like environment within Windows)
- Graphics Card (optional): For better parallel computation during model training (if using GPU-compatible libraries)

3.4 Software Requirements

The software requirements formulated for this project were:

- **Programming Languages:**
    - Python 3.10 or above
- **Operating system:**
    - Windows 11
- **Python Libraries:**
    - pandas — for data manipulation and CSV handling
    - numpy — for numerical operations
    - sklearn (Scikit-learn) — for ensemble machine learning models
    - nltk / spaCy — for Natural Language Processing
    - matplotlib, seaborn — for data visualization
    - jupyter notebook — for developing and running code interactively
- **Development Tools:**
    - VS Code/Jupyter lab/Anaconda

3.5 Summary

This chapter detailed the requirements for the mini-project, both functional and non-functional requirements, hardware specifications and software specifications. The requirements provided will ensure that the system can effectively generate synthetic EHR data, perform pre-processing, NLP analysis, train an ensemble model, and extract dark data. The detailed requirements provide a strong foundation for the implementation and evaluation of the proposed system. [13]

Chapter-4 Design of Mini Project

4.1 High Level Design

The high-level design provides an overview of the system's major components and how they interact to achieve the project goals. It outlines the key processes, starting from data generation to model training, prediction, and dark data extraction.

4.1.1 System Architecture

The system consists of four major components. To begin with, the Synthetic Data Generation module constructs a simulated EHR dataset for 500 patients. It contains 51 various attributes and is initially developed in JSON format, later transformed into CSV so that it can be utilized for analysis. Following this, the doctor's notes and other unstructured data like side effects undergo text pre-processing while numeric data undergoes cleaning, encoding, and normalization within the Pre-Processing module. By means of NLP, all the unstructured text data is transformed into a useful structured format with tokenization and stop word removal preceding the document's preparation. Subsequently, the ensemble machine learning models Random Forest and Gradient Boost, as well as NLP models, are incorporated into Model Training and NLP Analysis for classification, and useful pattern extraction from unstructured data, respectively. Through the use of trained models, the Dark Data Extraction and Visualization module uncovers concealed insights like survival probabilities, responses to treatment, and associated risks which aren't obvious in raw records. These insights are then depicted in plots and stored for further examination. [14]

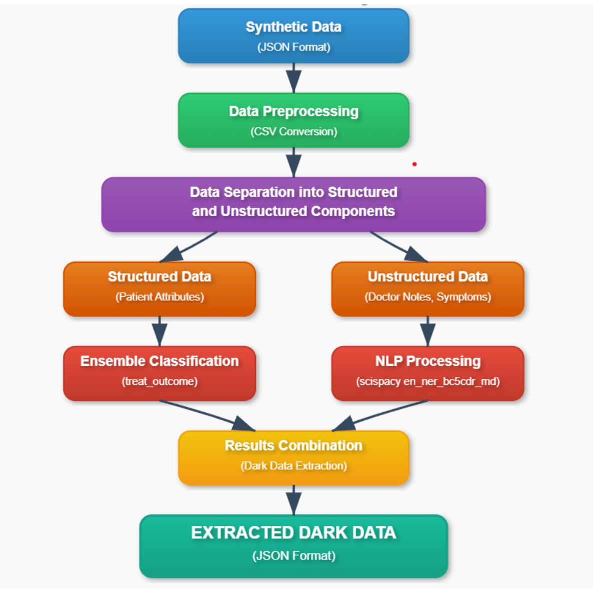


Fig 2. System Architecture

4.2 Detailed Design

The detailed design breaks down the project into modular components and discusses their structure, objectives, methodology, and expected outcomes.

4.2.1 Structure Chart

Electronic Health Records (EHR) commonly have massive amounts of data, including hidden, underutilized data, what we sometimes call dark data. This project intends to illustrate what hidden data can tell us using ensemble models and natural language processing.

- **Objective**: Create a system that can take the synthetic EHR data produced, process it, as well as apply ensemble learning models and incorporate the NLP techniques to help find hidden patterns lurking beneath the surface that may ultimately aid clinicians perform their own tasks and research.

- **Methodology**:

The project's model follows a structured approach: synthetic data was initially generated in JSON format and subsequently converted to CSV for further processing. Pre-processing steps were implemented to handle both structured and unstructured data effectively. Natural Language Processing (NLP) techniques were applied specifically to pre-process textual data, such as doctors' notes and other text-based information. Ensemble models, including Random Forests, Gradient Boosting, and Voting Classifiers, were employed for prediction and classification tasks. Additionally, visualizations and reporting played a crucial role in interpreting and understanding the insights extracted from the dark data. [15]

Expected outcome: The expectation for the system is that it will highlight patterns that aren't typically observed such as relationship between patient's history, and genetic markers, and the treatment they chose.

Conclusion: The project demonstrates the power of combining machine learning and natural language processing.

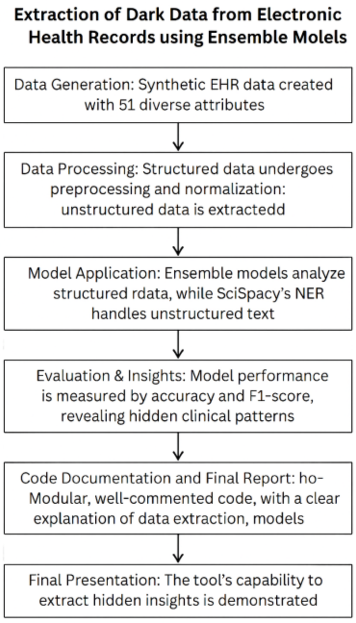

Extraction of Dark Data from Electronic Health Records using Ensemble Molels
Data Generation: Synthetic EHR data created with 51 diverse attributes
Data Processing: Structured data undergoes preprocessing and normalization: unstructured data is extractedd
Model Application: Ensemble models analyze structured rdata, while SciSpacy's NER handles unstructured text
Evaluation & Insights: Model performance is measured by accuracy and F1-score, revealing hidden clinical patterns
Code Documentation and Final Report: ho-Modular, well-commented code, with a clear explanation of data extraction, models
Final Presentation: The tool's capability to extract hidden insights is demonstrated

Fig 3. Structure Chart

4.2.2 Functional Description of the Modules

1. Data Generation Module

Input: Defined attributes and parameters for 500 synthetic patients in JSON format.

Output: A structured CSV file with 500 records and 51 attributes for further processing.

2. Pre-processing Module

The Pre-Processing Module operates by taking a CSV file as input, which contains both structured data (numerical and categorical) and unstructured text data. Its output is a cleaned and transformed dataset, fully prepared for model training. Key functions of this module include addressing missing values in the data, normalizing numerical attributes, encoding categorical variables, and performing text cleaning and tokenization for fields such as doctor's notes and side effects. This ensures the data is well-structured and suitable for downstream machine learning tasks.

3.NLP Analysis Module

The NLP Analysis Module takes textual data from fields such as doctor's notes and side effects columns as input and outputs processed textual features, including key symptoms, diagnosis trends, or medication effects. Its main functions include performing text tokenization and stop word removal to clean and structure the data effectively. It utilizes feature extraction techniques such as TF-IDF, Bag of Words, or word embeddings to derive meaningful patterns from the text. Additionally, the module conducts sentiment analysis or keyword-based analysis to uncover hidden insights, enabling further exploration of critical healthcare information.

4.Ensemble Model Training Module

The Ensemble Model Training Module is designed to take pre-processed structured data as input and produce trained ensemble models along with their predictions as output. Its key functions include training multiple machine learning models, such as Random Forest and Gradient Boosting, to leverage their combined strengths. The module employs ensemble techniques like majority voting or weighted averaging to integrate results from these models effectively. Additionally, it evaluates the performance of the models using metrics such as accuracy, precision, recall, and F1-score to ensure robust predictions, reliable outcomes. [16]

5. Dark Data Extraction Module

The Dark Data Extraction Module is designed to combine predictions from trained models and processed textual features to uncover hidden patterns, such as survival rates, relapse probabilities, or treatment side effects. Its core functions include merging structured model predictions with insights derived from NLP analysis, generating visualizations and detailed reports to present these hidden patterns effectively, and saving the final reports for further clinical or academic review. This module ensures that the extracted insights are accessible and actionable for deeper analysis or decision-making. [17]

Chapter-5 Implementation

This chapter discusses the actual implementation environment, tools, and conventions that were used in the execution of this mini-project. It discusses the platform and programming language choices and lists the libraries and tool usage during the development process. This section also discusses the coding conventions that were used to maintain the source code cleanly and usefully. [18]

5.1 Programming Language Selection

For this project, Python was selected as the primary programming language. Python is highly suitable for data science, machine learning, and natural language processing applications

5.2 Platform Selection

The platform chosen for this project is Google Colab, an online, cloud-based Jupyter notebook environment.

The libraries used are:

The following Python libraries were used:

- Pandas – for reading, writing, and processing structured data (CSV, JSON).
- NumPy – for numerical operations and array handling.
- Matplotlib & Seaborn – for data visualization.
- Scikit-learn – for building and evaluating ensemble models.
- NLTK – for basic NLP preprocessing (tokenization, stop-word removal).
- TfidfVectorizer (from scikit-learn) – for converting text data into numerical form.
- JSON – for handling JSON formatted data.
- OS – for file operations within directories.

5.3 Code Convention

To ensure clarity, readability, and efficiency, the following coding standards were adhered to: descriptive variable names were used to represent their specific purposes clearly; comments and documentation were provided for all functions and code blocks to enhance understanding. The code was organized into modular cells and functions to follow a reusable coding style. Regular indentation (4 spaces per level) was applied consistently for improved readability. Appropriate error messages and exception handling mechanisms were included where necessary, along with versioning of significant notebook milestones to track progress and prevent data loss. Additionally, the snake_case naming convention, as per Python PEP-8 guidelines, was utilized for both function names and variable names. [19]

5.4 Summary

This chapter discussed the implementation environment which comprises programming language, platform, libraries, and tools used in this project. Python was chosen because of its rich ecosystem, and ease of use; Google Colab was selected as the cloud-based development platform to use. This chapter also discussed the coding conventions that may have been used to ensure code clarity and efficiency for the entire process of development. [20]

Chapter-6 Experimental Results and Testing

This chapter provides an assessment of the system that was developed, through several stages of testing and performance assessment. It outlines the different metrics that were used to determine the system's effectiveness, the type of dataset employed in experimentation, and the results generated by the model that was implemented. In addition, the different levels of testing, namely unit testing, integration testing, and system testing have offered verification on whether the system components could produce a correct system, work together reliably, and formed a seamlessly integrated system. [21]

6.1 Evaluation Metrics

To measure the performance of the machine learning model developed for text extraction and classification from healthcare reports, the following parameters were employed:

- Accuracy: Number of instances classified correctly.
- Precision: Ratio of positive identifications that were indeed correct.
- Recall (Sensitivity): Total of true positives accurately identified.
- F1-Score: Harmonic mean of Precision and Recall, which gives an average performance measure.
- ROC-AUC Score: Measures the performance of the model in distinguishing between classes. [22]

6.2 Experimental Dataset

As part of this project, the data was self-generated. There were 500 simulated patient records, with each record denoting 51 fields. The dataset was initially in FHIR-standard JSON format, and then converted to CSV for easier manipulation, and to develop models more efficiently. The data contained structured fields (lab value, treatment, demographics) and unstructured fields (side effects, doctor comments) to simulate real-time typical health records.

6.3 Performance Analysis

In our project dealing with dark data extraction from Electronic Health Records (EHR) through ensemble learning methods, we performed comparative assessment of some ensemble classifiers to check their ability to detect latent or unused patterns. Among the models experimented, the Bagging Classifier recorded the best performance, with precision at 0.83, recall at 0.82, F1-score at 0.82, and accuracy at 0.82, testifying to its strong ability to consistently identify useful data points with few false positives or false negatives. XGBoost and Stacking Classifier were also top performers, recording balanced F1-scores of 0.72 and 0.70 respectively, a testament to their strengths in recognizing complex patterns in the data. Conversely, the Random Forest model performed the worst with an F1-score of 0.45, indicating that it might not be the best for this particular task. In general, the findings highlight the benefits of sophisticated ensemble techniques, including bagging and boosting strategies, in revealing dark data within EHR systems and enhancing data-driven clinical findings.

Table 1: Performance Analysis Summary Table

Classifier	Precision	Recall	F1-Score	Accuracy
Random Forest	0.46	0.47	0.45	0.47
Bagging Classifier	0.83	0.82	0.82	0.82
Gradient Boosting	0.69	0.68	0.68	0.68
XGBoost	0.72	0.72	0.72	0.72
LightGBM	0.69	0.68	0.68	0.68
Voting Classifier (Hard)	0.69	0.69	0.68	0.69
Stacking Classifier	0.71	0.70	0.70	0.70

6.4 Unit Testing

The focus of unit testing has always been verifying the correct functioning of the system's components, examining in detail the JSON to CSV parsing subroutine, text processing procedures like tokenization, stopword elimination, and the TF-IDF vectorization, model training, and prediction processes. These activities were done for each component using different sample data to validate their treatment of missing values, text cleaning and preprocessing, format conversions, control results, and output results. Such a rigorous practice made sure that the system checked every part with precision.

6.5 Integration Testing

For integration testing, the unit functions were combined and tested as one system to verify the seamless interaction among the various subsystems, in particular the incorporation of the processed CSV file, the JSON derived one, into the preprocessing phase of the TF-IDF vectorization and model prediction, and also if the results were saved and retrieved as required in CSV and JSON formats through the correct pathways and locations before being made accessible later.

6.6 System Testing

System testing was conducted to ensure the overall functionality of the application. During this process, the system executed data ingestion through file upload in JSON or CSV format, text cleaning and extraction thereafter, conversion of cleaned data into a suitable template, model execution with output storage, and prediction. Throughout multiple test iterations, the system was stable and consistent when dealing with various input formats and providing correct and well-structured results.

6.7 Summary

This chapter discusses the experimental results and testing procedures to assess the performance and dependability of the implemented system. Both classification models such as Random Forest, Bagging, Gradient Boosting, XGBoost, LightGBM, Voting Classifier, and Stacking Classifier were examined and compared with evaluation metrics such as precision, recall, F1-Score, and accuracy. Bagging Classifier yielded the highest overall accuracy metric compared to the XGBoost and other ensemble models, which performed reasonably well. [23]

This chapter also summarizes the multiple levels of testing on the system, to showcase the system's correctness and integration abilities. Unit Testing was performed to verify a specific module, Integration Testing verified the interaction between components and data consistency for a smooth data flow, and System Testing was performed to verify the functionality, dependability, and performance of the complete and integrated system based on the defined requirements.

This well-structured evaluation has shown the strengths and weaknesses of the system and has provided a strong start for future improvements. [24]

Chapter-7 Conclusion and Future Enhancement

The project developed a complete process and assessment systems to process and analyze synthetic electronic health records for blood cancer patients that utilize both structured and unstructured data. Using a self-generated, FHIR-compliant JSON dataset of blood cancer patients, the research converted the dataset from JSON to CSV, applied data preparation, took a driven approach to text data extraction and classification based on machine learning, and results showed promising performance measures, allowing us to conclude that this type of system can successfully analyze complex types of datasets relevant to the practice of medicine. The assessment process could be enhanced in the future by complimenting it with actual clinical data, deploying the assessment process to a web-based application, creating a more complete model driven from the synthetic data, and by enhancing our capabilities in natural language processing (NLP) to aid in extracting insights from text and/or by implementing named entity recognition (NER) techniques. In addition, expanding the functionality of the model to support multi-disease models would provide further application to the healthcare field. [25]

References

[1] C.-M. Hsu, L.-L. Liang, Y.-T. Chang and W.-C. Juang, “Emergency department overcrowding: Quality improvement in a Taiwan Medical Center,” *Journal of the Formosan Medical Association,* vol. 118, p. 186–193, 2019.

[2] M. Viterbo, “Estratégias de gestão para redução da aglomeração e superlotação no pronto socorro adulto de um hospital terciário da zona norte da cidade de São Paulo-SP,” Gestão pronto socorro, 2020.

[3] F.-Y. Cheng and e. al., “Using machine learning to predict ICU transfer in hospitalized COVID-19 patients,” *Journal of Clinical Medicine,* vol. 9, p. 1668, 2020.

[4] G. Lombardo, C. Couvert, M. Kose, A. Begum and C. Spiertz, “Electronic health records (EHRs) in clinical research and platform trials: Application of the innovative EHR-based methods developed by EU-PEARL,” *Journal of Biomedical Informatics,* vol. 148, no. 104553, pp. 1-8, 2023.

[5] Y. Ramakrishnaiah, N. Macesic, G. I. Webb, A. Y. Peleg and S. Tyagi, “EHR-QC: A streamlined pipeline for automated electronic health records standardisation and preprocessing to predict clinical outcomes,” *Journal of Biomedical Informatics,* vol. 147, no. 104509, p. 1–10, 2023.

[6] H. Niu, O. A. Omitaomu, M. A. Langston, M. Olama, O. Ozmen, H. B. Klasky, A. Laurio, M. Ward and J. Nebeker, “EHR-BERT: A BERT-based model for effective anomaly detection in electronic health records,” *Journal of Biomedical Informatics,* vol. 150, no. 104605, p. 1–11, 2024.

[7] M. Herrero-Zazo, T. Fitzgerald, V. Taylor, H. Street, A. N. Chaudhry, J. R. Bradley, E. Birney and V. L. Keevil, “Using machine learning to model older adult inpatient trajectories from electronic health records data,” *iScience,* vol. 28, no. 105876, p. 1–15, 2022.

[8] H. Atasoy, B. N. Greenwood and J. S. McCullough, “The digitization of patient care: a review of the effects of electronic health records on health care quality and utilization,” *Annual Review of Public Health,* vol. 40, no. 203, p. 487–500, 2019.

[9] S. Džeroski and B. Ženko, “Is combining classifiers with stacking better than selecting the best one?,” *Machine Learning,* vol. 54, p. 255–273, 2004.

[10] A. S. Sarvestani, A. A. Safavi, N. M. Parandeh and M. Salehi, “Predicting breast cancer survivability using data mining techniques,” in *2nd International Conference on Software Technology and Engineering*, Bengaluru, 2010.

[11] M. M. Mehdy, P. Y. Ng, E. F. Shair, N. I. Saleh and C. Gomes, “Artificial neural networks in image processing for early detection of breast cancer,” *ScienceAI,* vol. 23, no. 12343, pp. 20-40, 2020.

[12] L. Breiman, J. Friedman, C. J. Stone and R. A. Olshen, Classification and Regression Trees, 1984: CRC Press, New York.

[13] Y. Freund, R. Schapire and N. Abe, “A short introduction to boosting,” *Journal of the Japanese Society for Artificial Intelligence,* vol. 14, p. 612, 1999.

[14] F. Rustam and e. al., “Classification of Shopify app user reviews using novel multi text features,” *IEEE Access,* vol. 8, p. 30234–30244, 2020.

[15] M. R. Anderson, D. Antenucci, V. Bittorf, M. Burgess, M. J. Cafarella, A. Kumar, F. Niu, Y. Park, C. Re and C. Zhang, “A data system for feature engineering,” in *CIDR 2013, Sixth Biennial Conference on Innovative Data Systems Research*, 2013.

[16] B. Hurst, B. Tekinerdogan, T. Alskaif, A. Boddy and N. Shone, “Securing electronic health records against insider-threats: A supervised machine learning approach,” *Smart Health,* vol. 26, no. 100354, p. 1–14, 2022.

[17] M. J. Wainwright and M. I. Jordan, “Log-determinant relaxation for approximate inference in discrete Markov random fields,” *IEEE Transactions on Signal Processing,* vol. 54, no. 6–1, p. 2099–2109, 2006.

[18] I. o. M. (. C. o. D. S. f. P. Safety, “Key Capabilities of an Electronic Health Record System: Letter Report,” National Academies Press (US), Washington, DC, 2003.

[19] S. M. Powsner, J. C. Wyatt and P. Wright, “Opportunities for and challenges of computerisation,” *The Lancet,* vol. 352, p. 1617–1622, 1998.

[20] C. A. Weaver and P. Teenier, “ Rapid EHR development and implementation using web and cloud-based architecture in a large home health and hospice organization,” *Studies in Health Technology and Informatics,* vol. 201, p. 380–387, 2014.

[21] M. Rafało, “Cross validation methods: Analysis based on diagnostics of thyroid cancer metastasis,” *ICT Express,* vol. 23, p. 183–188, 2022.

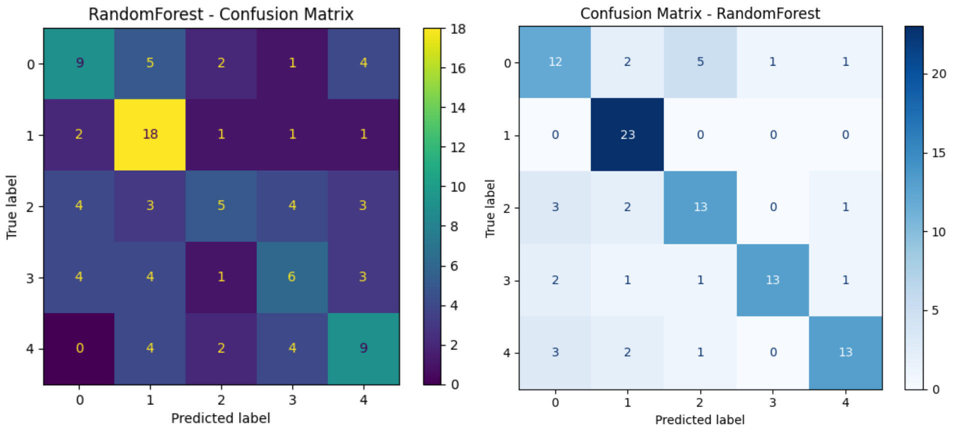
[22] V. Jaiswal, P. Saurabh, U. K. Lilhore, M. Pathak, S. Simaiya and S. Dalal, “A breast cancer risk prediction and classification model with ensemble learning and big data fusion,” *Decision Analytics Journal,* vol. 8, no. 100298, p. 1–13, 2023.

[23] A. Srinivas and J. P. Mosiganti, “A brain stroke detection model using soft voting based ensemble machine learning classifier,” *Measurement: Sensors,* vol. 29, no. 100871, pp. 1-7, 2023.

[24] H. Horng, J. Steinkamp, C. Kahn Jr. and T. Cook, “Ensemble Approaches to Recognize Protected Health Information in Radiology Reports,” *Journal of Digital Imaging,* vol. 35, no. 6, p. 1694–1698, 2022.

[25] C. Stevens, A. Lyons, K. Dharmayat, A. Mahani, K. Ray, A. Vallejo-Vaz and M. Sharabiani, “Ensemble Machine Learning Methods in Screening Electronic Health Records: A Scoping Review,” *Digital Health,* vol. 9, no. 11, p. e1002695, 2023.

Appendices



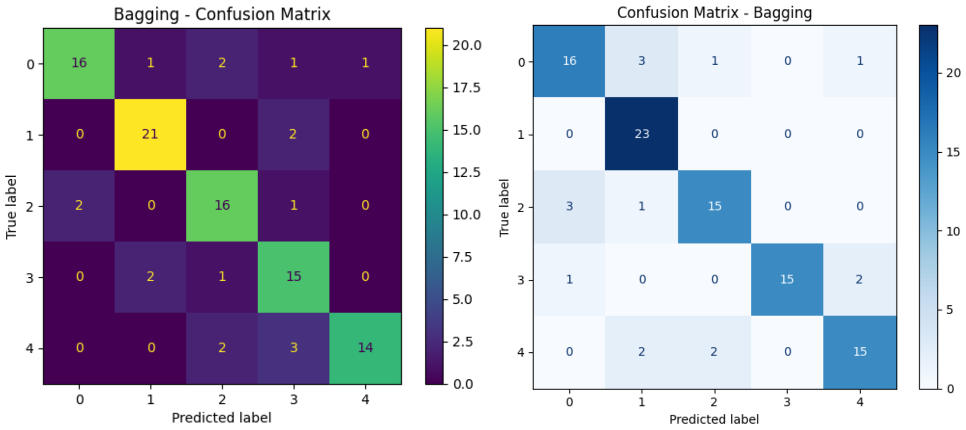
RandomForest - Confusion Matrix

True label \ Predicted label	0	1	2	3	4
0	9	5	2	1	4
1	2	18	1	1	1
2	4	3	5	4	3
3	4	4	1	6	3
4	0	4	2	4	9

Confusion Matrix - RandomForest

True label \ Predicted label	0	1	2	3	4
0	12	2	5	1	1
1	0	23	0	0	0
2	3	2	13	0	1
3	2	1	1	13	1
4	3	2	1	0	13

Figure 4: Confusion Matrix Comparison of Random Forest Classifier with and Without Dark Data



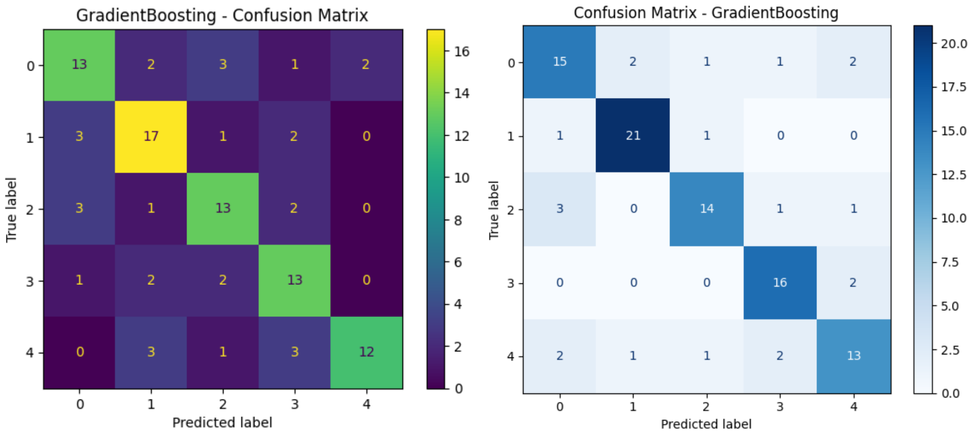
Bagging - Confusion Matrix

True label \ Predicted label	0	1	2	3	4
0	16	1	2	1	1
1	0	21	0	2	0
2	2	0	16	1	0
3	0	2	1	15	0
4	0	0	2	3	14

Confusion Matrix - Bagging

True label \ Predicted label	0	1	2	3	4
0	16	3	1	0	1
1	0	23	0	0	0
2	3	1	15	0	0
3	1	0	0	15	2
4	0	2	2	0	15

Figure 5: Confusion Matrix Comparison of Bagging Classifier with and Without Dark Data



GradientBoosting - Confusion Matrix

True label \ Predicted label	0	1	2	3	4
0	13	2	3	1	2
1	3	17	1	2	0
2	3	1	13	2	0
3	1	2	2	13	0
4	0	3	1	3	12

Confusion Matrix - GradientBoosting

True label \ Predicted label	0	1	2	3	4
0	15	2	1	1	2
1	1	21	1	0	0
2	3	0	14	1	1
3	0	0	0	16	2
4	2	1	1	2	13

Figure 6: Confusion Matrix Comparison of Gradient Boosting with and Without Dark Data

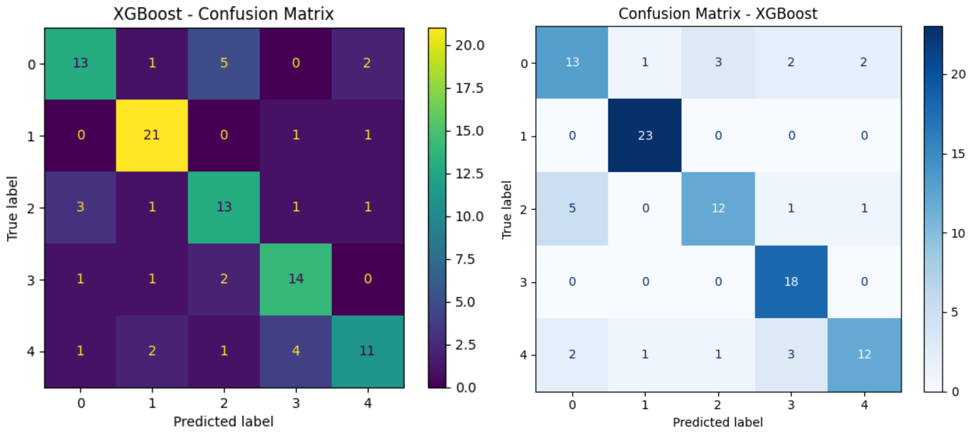


Figure 7: Confusion Matrix Comparison of XGBoost Classifier with and Without Dark Data

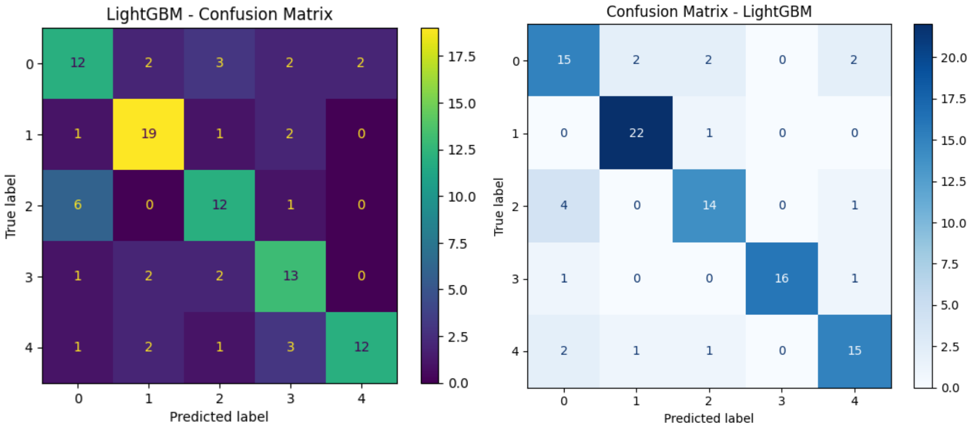


Figure 8: Confusion Matrix Comparison of LightGBM Classifier with and Without Dark Data

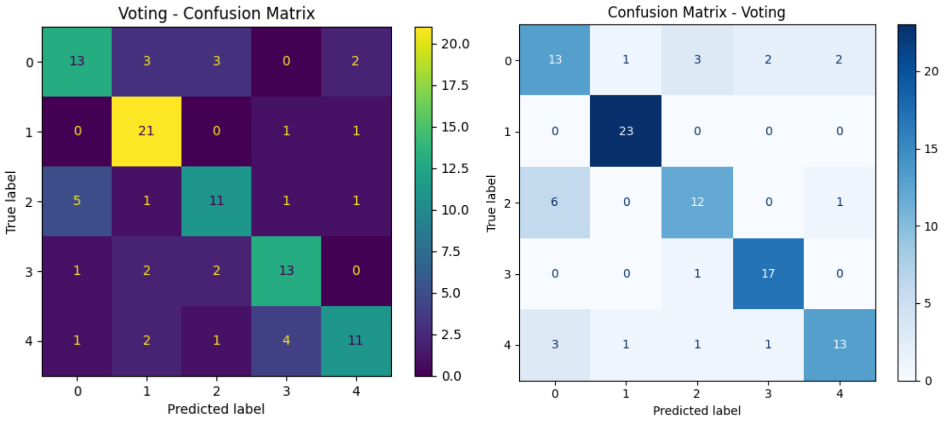


Figure 9: Confusion Matrix Comparison of Voting Classifier with and Without Dark Data

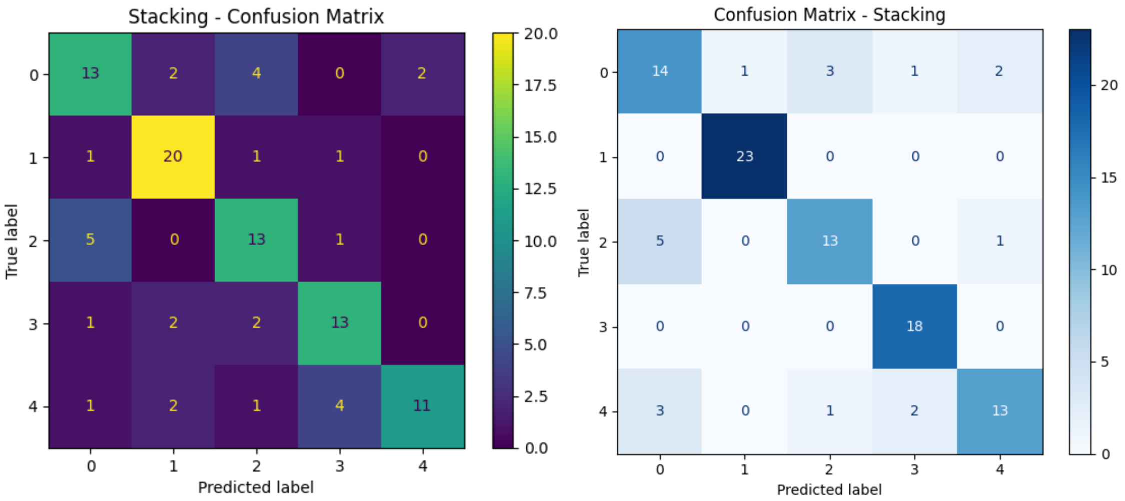


Figure 10: Confusion Matrix Comparison of Stacking Classifier with and Without Dark Data

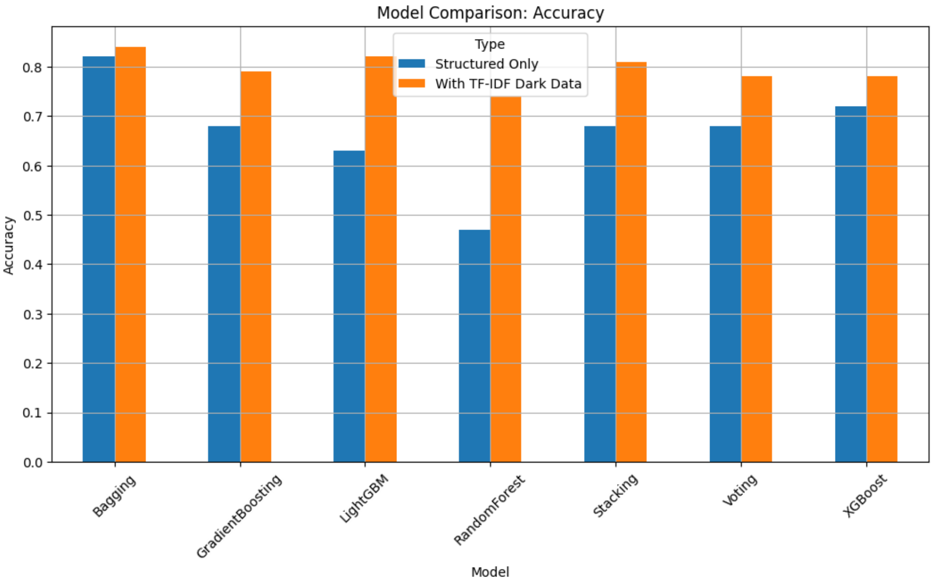


Figure 11: Model Comparison: Accuracy

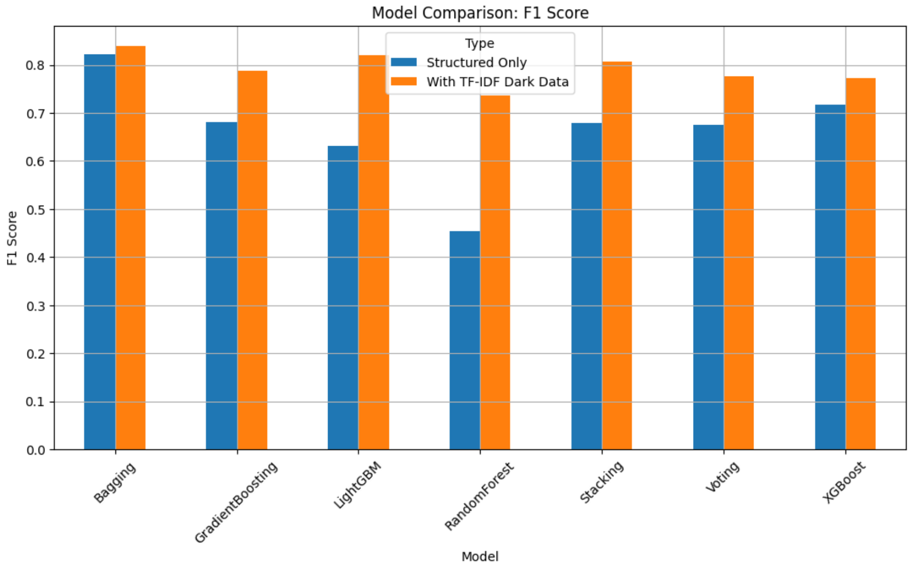
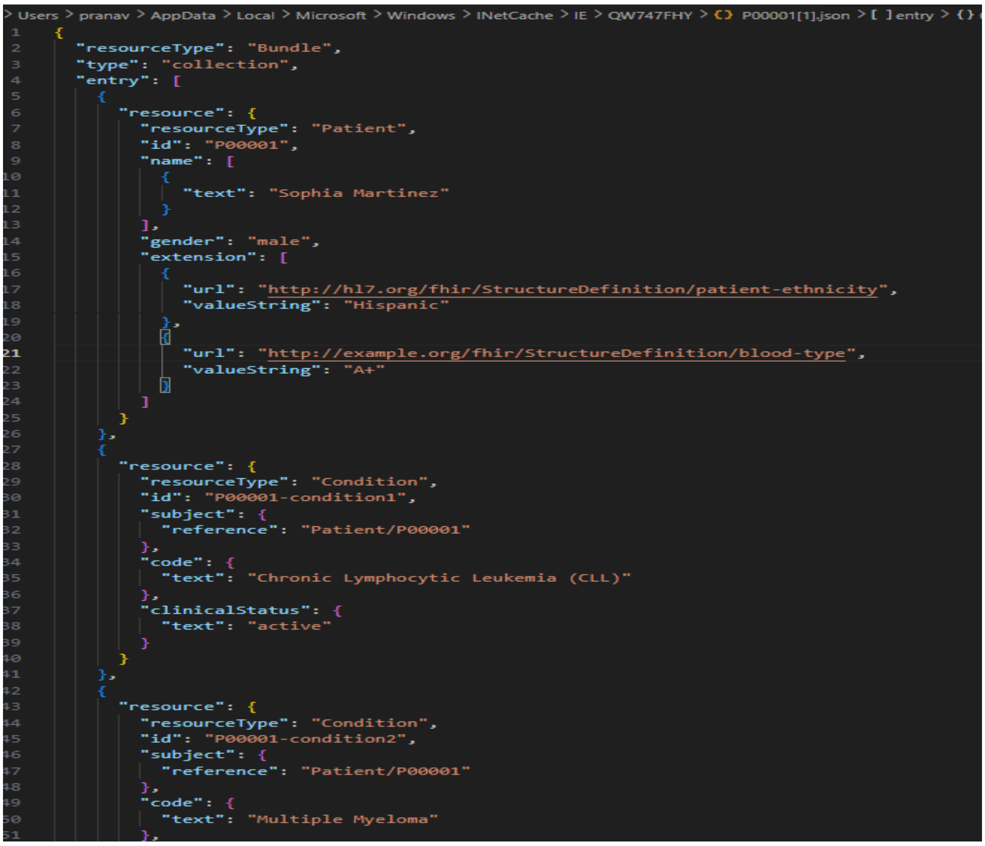


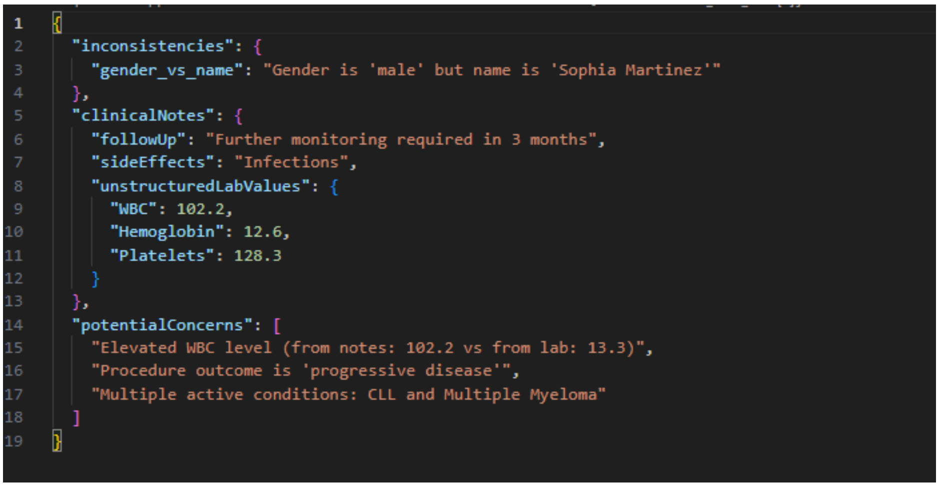
Figure 12: Model Comparison: F1 Score



> Users > pranav > AppData > Local > Microsoft > Windows > INetCache > IE > QW747FHY > {} P00001[1].json > [ ] entry > {}

```
{
  "resourceType": "Bundle",
  "type": "collection",
  "entry": [
    {
      "resource": {
        "resourceType": "Patient",
        "id": "P00001",
        "name": [
          {
            "text": "Sophia Martinez"
          }
        ],
        "gender": "male",
        "extension": [
          {
            "url": "http://hl7.org/fhir/StructureDefinition/patient-ethnicity",
            "valueString": "Hispanic"
          },
          {
            "url": "http://example.org/fhir/StructureDefinition/blood-type",
            "valueString": "A+"
          }
        ]
      }
    },
    {
      "resource": {
        "resourceType": "Condition",
        "id": "P00001-condition1",
        "subject": {
          "reference": "Patient/P00001"
        },
        "code": {
          "text": "Chronic Lymphocytic Leukemia (CLL)"
        },
        "clinicalStatus": {
          "text": "active"
        }
      }
    },
    {
      "resource": {
        "resourceType": "Condition",
        "id": "P00001-condition2",
        "subject": {
          "reference": "Patient/P00001"
        },
        "code": {
          "text": "Multiple Myeloma"
        },
```

Figure 13: Synthetic blood cancer EHR data



```
{
  "inconsistencies": {
    "gender_vs_name": "Gender is 'male' but name is 'Sophia Martinez'"
  },
  "clinicalNotes": {
    "followUp": "Further monitoring required in 3 months",
    "sideEffects": "Infections",
    "unstructuredLabValues": {
      "WBC": 102.2,
      "Hemoglobin": 12.6,
      "Platelets": 128.3
    }
  },
  "potentialConcerns": [
    "Elevated WBC level (from notes: 102.2 vs from lab: 13.3)",
    "Procedure outcome is 'progressive disease'",
    "Multiple active conditions: CLL and Multiple Myeloma"
  ]
}
```

Figure 14: Extracted dark data from Synthetic EHR data

pranav kalp

Final Report-2.docx

My Files

My Files

RV University

Document Details

Submission ID

trn:oid:::18459:93266443

Submission Date

Apr 28, 2025, 12:32 PM GMT+5:30

Download Date

Apr 28, 2025, 12:35 PM GMT+5:30

File Name

Final Report-2.docx

File Size

891.2 KB

30 Pages

5,015 Words

30,503 Characters

13% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Match Groups

69 Not Cited or Quoted 13%
Matches with neither in-text citation nor quotation marks

0 Missing Quotations 0%
Matches that are still very similar to source material

2 Missing Citation 0%
Matches that have quotation marks, but no in-text citation

0 Cited and Quoted 0%
Matches with in-text citation present, but no quotation marks

Top Sources

6%	Internet sources
3%	Publications
11%	Submitted works (Student Papers)

Integrity Flags

0 Integrity Flags for Review

No suspicious text manipulations found.

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.

Match Groups

69 Not Cited or Quoted 13%
Matches with neither in-text citation nor quotation marks

0 Missing Quotations 0%
Matches that are still very similar to source material

2 Missing Citation 0%
Matches that have quotation marks, but no in-text citation

0 Cited and Quoted 0%
Matches with in-text citation present, but no quotation marks

Top Sources

6%	Internet sources
3%	Publications
11%	Submitted works (Student Papers)

Top Sources

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1	Submitted works	**Liverpool John Moores University on 2024-08-30**	**<1%**
2	Submitted works	**Turun yliopisto on 2021-06-17**	**<1%**
3	Internet	**ntnuopen.ntnu.no**	**<1%**
4	Submitted works	**National Institute of Technology, Rourkela on 2021-07-26**	**<1%**
5	Submitted works	**Arab Open University on 2024-11-07**	**<1%**
6	Submitted works	**Vardhaman College of Engineering, Hyderabad on 2022-05-10**	**<1%**
7	Internet	**1library.org**	**<1%**
8	Submitted works	**Middlesex University on 2008-06-16**	**<1%**
9	Internet	**propulsiontechjournal.com**	**<1%**
10	Submitted works	**Liverpool John Moores University on 2025-04-12**	**<1%**

11	Submitted works	kitsw on 2025-04-13	<1%
12	Submitted works	Blue Mountain Hotel School on 2024-12-04	<1%
13	Submitted works	University of Warwick on 2022-06-26	<1%
14	Publication	Altarturi, Hamza H. M.. "Cyber Parental Control Framework for Objectionable We...	<1%
15	Submitted works	Arab Open University on 2024-11-07	<1%
16	Submitted works	Victoria University on 2024-06-12	<1%
17	Internet	disertasipascasarjanauniversitas.blogspot.com	<1%
18	Submitted works	Nottingham Trent University on 2025-04-25	<1%
19	Submitted works	Higher Education Commission Pakistan on 2015-05-22	<1%
20	Submitted works	Queen's University of Belfast on 2025-04-14	<1%
21	Internet	dr.ntu.edu.sg	<1%
22	Internet	fastercapital.com	<1%
23	Internet	www.geeksforgeeks.org	<1%
24	Submitted works	Napier University on 2024-12-11	<1%

25	Submitted works	University of East London on 2012-12-05	<1%
26	Internet	irjet.net	<1%
27	Submitted works	RMIT University on 2023-08-27	<1%
28	Internet	ir.uitm.edu.my	<1%
29	Internet	researchnetwork.net	<1%
30	Internet	www.ijettjournal.org	<1%
31	Submitted works	BB9.1 PROD on 2025-04-22	<1%
32	Submitted works	UCL on 2025-04-25	<1%
33	Internet	arxiv.org	<1%
34	Internet	www.knoxnews.com	<1%
35	Submitted works	Liverpool John Moores University on 2024-12-14	<1%
36	Submitted works	Liverpool John Moores University on 2025-01-06	<1%
37	Publication	R. N. V. Jagan Mohan, B. H. V. S. Rama Krishnam Raju, V. Chandra Sekhar, T. V. K. P...	<1%
38	Submitted works	University of Bradford on 2025-04-14	<1%

39 Submitted works

University of Westminster on 2025-04-15 <1%

40 Publication

Yuh-Chuan Shih, Sheau-Farn Max Liang, Yu-Hsing Huang, Yu-Cheng Lin, Chih-Lon... <1%

41 Internet

centaur.reading.ac.uk <1%

42 Internet

export.arxiv.org <1%

43 Internet

philarchive.org <1%

44 Internet

sode-edu.in <1%

45 Submitted works

ESoft Metro Campus, Sri Lanka on 2025-02-15 <1%

46 Submitted works

Higher Education Commission Pakistan on 2019-04-23 <1%

47 Publication

"Agents and Multi-agent Systems: Technologies and Applications 2024", Springer ... <1%

48 Submitted works

Central Queensland University on 2024-08-23 <1%

49 Publication

H L Gururaj, Francesco Flammini, V Ravi Kumar, N S Prema. "Recent Trends in He... <1%

50 Submitted works

University of Westminster on 2025-04-07 <1%