Proceedings of the International Conference on Sustainable Communication Networks and Application (ICSCN-2025)
DVD Part Number: CFP25DW8-DVD; ISBN: 979-8-3315-9419-0

# Healthcare Analytics: Data Mining Techniques on Electronic Health Records (EHR) to Build a Predictive Model for Early Disease Diagnosis

Siddique Ibrahim S. P.
Assistant Professor,
School of Computer Science and Engineering,
VIT-AP University, Amaravati,
Andhra Pradesh, India
siddiqueibrahim.sp.cse@kct.ac.in

Abhik Das
School of Computer Science and Engineering,
VIT-AP University, Amaravati,
Andhra Pradesh, India
abhikdas0811@gmail.com

Nishant Gaurav
School of Computer Science and Engineering,
VIT-AP University, Amaravati,
Andhra Pradesh, India
nishantgaurav2208@gmail.com

Laxmi Prasad Mishra
School of Computer Science and Engineering,
VIT-AP University, Amaravati,
Andhra Pradesh, India
laxmiprasadmishra04@gmail.com

Priyanshu Kanyal
School of Computer Science and Engineering,
VIT-AP University, Amaravati,
Andhra Pradesh, India
kanyalpriyanshu1@gmail.com

***Abstract*— This paper examines how various machine learning algorithms and data mining methods can be used on Electronic Health Records (EHRs) to diagnose diseases early. It examines the challenges of using different machine learning algorithms for predicting different diseases, which can reduce efficiency in medical practice. To address these problems, a Hybrid Disease Aware Ensemble Model is proposed. The novelty of this framework lies in its use of disease-specific weights and calibrated thresholds, which is a departure from the traditional use of a single global meta learner. This approach balances sensitivity and specificity, providing a robust and unified machine learning framework for predicting multiple diseases.**

***Keywords*— *Electronic Health Records (EHR), Machine Learning, Predictive Modeling, Early Disease Diagnosis, Missing Data***

## I. INTRODUCTION

Early disease diagnosis is a necessity to reduce the risk of aggravated medical conditions and reducing healthcare expenditure. Electronic Healthcare Records (EHRs) consist of clinical, demographic, and laboratory data, which can help in predictive analytics. This is done by utilizing machine learning algorithms, which can identify correlations in the data.

However, the majority of EHR data is used for clinical and administrative purposes, which is not suitable for predictive analysis. It might contain missing values, unstructured medical documents, which hinder the implementation. Data privacy concerns have also made it difficult for widespread use. Nevertheless, proper use of EHRs can help build a predictive model.

Machine learning models are implemented on EHRs to build a disease prediction model [1]. This helps in early diagnosis of disease, potentially saving a patient's life and reducing medical costs. The challenges that exist with these predictive models are that there are different machine learning algorithms used to predict different diseases [2]. While the results are satisfactory, it makes it difficult to trust which algorithm to use. Since there is no fixed framework due to different machine learning models used, medical professionals find it difficult to choose which model to rely on to diagnose a disease.

To address this limitation, this study proposes a Hybrid Disease-Aware Ensemble model that combines multiple machine learning algorithms with disease-specific weighting and calibrated thresholds. This approach helps to provide a unified framework for predicting different diseases, balancing sensitivity, and specificity to reduce false positives.

### A. Our Contributions

The key contributions of this paper are as follows:

1. A Novel Hybrid Ensemble Model: We propose a Hybrid Disease-Aware Ensemble model that, unlike traditional stacked ensembles, utilizes disease-specific weights and thresholds to improve predictive accuracy and clinical relevance.
2. Balanced Sensitivity and Specificity: Our model introduces a threshold-tuning mechanism that enforces a minimum specificity, directly addressing the common problem of high false-positive rates in disease prediction.
3. A Unified Predictive Framework: We demonstrate a single, robust framework capable of predicting multiple distinct diseases, overcoming the limitations of multiple models for multiple diseases in existing literature.

## II. LITERATURE SURVEY

### A. Data Preprocessing and Data Cleaning

Before we can proceed to applying machine learning algorithms to Electronic Health Records (EHRs), raw, unpolished data must be converted to a structured format that

can be easily analysed. Inconsistent health records of different patients, different writing structures used in various medical institutions, result in "untidy" data that must be refined.

The first step of the steps involved in Data Preprocessing is dealing with missing data. Missing data can happen due to improper documentation. A patient may not show a particular symptom of a disease. Instead of entering a value under the specific symptom, it is left blank, that results in missing data. Simple strategies exist to handle missing data, which is dropping the rows containing missing data, but this can risk losing vital data from other variables.

Instead of dropping rows, imputation techniques are used. Mean, Median, or Mode imputation is a basic approach to replace missing data with the mean, median and mode of the data [24]. This technique is easy and fast but it can distort the variance and correlations in the data. A more robust approach would be to use Multiple Imputation by Chained Equations (MICE) [4]. It creates complete datasets by modelling each feature with missing values as a function of other variables. This preserves the correlations but is computationally expensive. A sophisticated strategy is Hybrid Imputation, which uses the K-Nearest Neighbours (KNN) Algorithm, providing better data reliability [3].

### *B. Feature Engineering*

After Data Cleaning, the datasets can be enhanced further by creating meaningful new attributes from existing data. This can even involve extracting temporal features from timestamp data. For example, for regularly visiting patients, the average clinical visits each month can be tracked to understand if there is any seasonal factor in the illness of the patient that leads to frequent or seldom medical appointments. Several feature engineering techniques exist which can be highly effective:

1. Composite Scores: Multiple raw features can be combined to create a medically relevant feature. For example, the Body Mass Index (BMI) of a patient can be calculated from the height and weight of the patient.
2. Binning: Continuous data can be grouped into categories. For example, a patient's age can be categorized into 'Child,' 'Adult' and 'Senior' [16].

### *C. Comparative Analysis of Predictive Models*

#### *1) Single Task Learning (STL) Models*

The least complex statistical model approach is Logistics Regression. It is simple to comprehend, yet unable to handle complex and non-linear relations present in EHR data. It offers moderate performance with low sensitivity. An extension of Logistic Regression is Support Vector Machines (SVM), Random Forests, and Gradient Boosting. Heart Failure can be predicted using Support Vector Machines [24]. Risks of diabetes can be predicted from EHRs using Random Forests algorithm [25]. Gradient Boosting has also been used to recommend drugs through personalised drug information [26]. These STL models perform well with high accuracy for individual tasks.

It fails to address the problem of comorbidities, which requires modelling the relationship between diseases.

#### *2) Ensemble Learning*

To produce a more robust predictive model, Ensemble Learning techniques are utilized, which combine multiple weak or base learners [6]. It comprises the following strategies:

1. Bagging as embodied through Random Forest aims at variance reduction through training hundreds of models atop numerous data subsets and then combining their outputs [23].
2. Boosting (e.g., XGBoost, AdaBoost, Gradient Boosting): It trains models one at a time. Each model aims at correcting mistakes of the prior model [23].

Ensemble learning utilizes stacking. It is an influential ensemble method in which outputs of several base learners are utilized as input features of a meta-model. The stacked ensemble technique outperforms any single model [23]. However, the conventional approach uses a single global meta-learner that learns a fixed set of weights for combining base models, treating predictions for all diseases uniformly. This approach may not always work as the optimal base model may differ from one disease to another.

Table I. below shows the different machine learning models used for different diseases supported with its Key Performance Metric.

### *D. Research Gaps and Motivation*

The literature review reveals two major gaps in existing research. First, the Single Task Learning (STL) models, while effective for individual diseases, neglect comorbidities, which are prevalent in clinical scenarios. Second, the traditional ensemble models usually adopt a single global meta-learner, treating all diseases uniformly. This approach does not take into consideration that the predictability of base models might differ significantly across different diseases. This absence of a single, disease-adaptive framework creates a considerable challenge for medical practitioners seeking a reliable diagnostic tool. Our motivation stems from closing this gap by developing a hybrid model that remains attentive to the unique characteristics of each disease. Table II. Demonstrates the comparison between the existing machine learning methodologies and the proposed machine learning model.

TABLE I.: DIFFERENT MACHINE LEARNING MODELS USED FOR DIFFERENT DISEASES WITH KEY PERFORMANCE METRIC

| Machine Learning Model | Disease | Key Performance Metric | Citation |
|---|---|---|---|
| Logistic Regression | Heart Failure | AUROC = 0.79, Sensitivity = 0.75, Precision = 0.70 | [14] |
| Random Forest | Heart Disease | AUROC = 0.95, Accuracy = 0.92 | [19] |
| Logistic Regression | Sepsis | AUROC = 0.952, Specificity = 0.87 to 0.93, Sensitivity = 0.878 to 0.922 | [14] |
| Decision Tree | Type-2 Diabetes | AUROC = 0.948, Accuracy = 0.94, Precision = 0.94 | [15] |
| ANN + Logistic Regression | Liver Cancer | AUROC = 0.873, Specificity = 0.755, Specificity = 0.757 | [20] |
| Random Forest | Gastric Cancer | AUROC = 0.75, F-measure = 0.71 | [21] |

TABLE II.: COMPARISION BETWEEN EXISTING METHODOLOGIES AND THE PROPOSED METHODOLOGY

| Approach | Methodology | Key Limitations | How our Proposed Model Differs |
|---|---|---|---|
| STL Models (e.g., [14], [15]) | Logistic Regression, Decision Tree | Single-disease focus; fails to address comorbidities. | Provides a unified framework for multiple diseases. |
| Conventional Ensemble (Stacking) | Global meta-learner combines base models. | Treats all diseases uniformly; may not be optimal for diseases where base models perform differently. | Introduces disease-specific weights to leverage the best-performing algorithm for each disease. |
| Proposed Hybrid Model | Disease-Aware Ensemble with specific weights and thresholds. | N/A | Balances sensitivity and specificity through per-disease threshold tuning; improves clinical reliability. |

## III. PROPOSED SYSTEM METHODOLOGY

To address the problem of multiple machine learning models being used for different diseases, this research is proposed in two stages. We start with a Stacked Ensemble Model as a baseline and then move on to a Hybrid Disease-Aware Ensemble Model, which better addresses the challenges of an early disease prediction.

### *A. Dataset Description*

The dataset utilized for this methodology is a composite collection, aggregated from several publicly available, widely used medical datasets sourced from the Kaggle platform. This approach was chosen to create a heterogenous dataset that tests the proposed model's ability to predict multiple diseases.

The following datasets have been used:

1. Cancer: The Benign and Malignant Cancer Data was used. This dataset consists of 137 samples across 31 features [8].
2. Breast Cancer: The Breast Cancer Wisconsin (Diagnostic) Dataset was used. This dataset consists of 136 samples across 10 features [9].
3. Lung Cancer: The Survey Lung Cancer dataset was used. Contains 284 instances with 16 features [10].
4. Diabetes: The Pima Indians Diabetes Database was included. It consists of 149 samples across 8 features [11].
5. Heart Disease: The Cleveland Clinic Heart Disease Dataset. It consists of 303 observations over 14 features [12].
6. Chronic Kidney Disease (CKD): This dataset contains 400 instances and 25 features [13].

### *B. Data Preparation*

The dataset consists of patients' health records with various features, which include demographic information, clinical measurements, lab results, and disease labels.

Feature Identification: Features have been divided into numerical and categorical types.

Preprocessing:

1. Missing values of numerical variables are imputed using K-Nearest Neighbours (KNN).
2. Numerical features are also standardized using StandardScaler.
3. Those discrete features like the disease label are converted into a numerical one through One-Hot Encoding.

### C. Handling Class Imbalance

Synthetic Minority Oversampling Technique (SMOTE) has also been utilized in solving class imbalancing [5]. It generated synthetic copies of minority classes and prevents base models from being biased towards majority classes.

### D. Base Learners

Two machine learning algorithms have been used as base learners

1. Random Forest Algorithm (RF): Via training of many models on different subsets of the dataset and then averaging their predictions, this method compensates for variance.
2. XGBoost Algorithm (XGB): The approach develops the models sequentially such that each next model works towards the elimination of the flaws produced by the precursors.

### E. Stacked Ensemble (Baseline)

Alongside the base models, we introduce a meta-model called Logistic Regression that gets trained with the base learners' outputs to make the final prediction. The model obtains global weights intended to optimize the final decision boundary irrespective of variation-specific to diseases.

Stacked ensemble uses base learners' probabilities as input to a meta-learner (logistic regression).

$$P_{stacked} = \sigma\left(\beta_0 + \beta_1 P_{RF} + \beta_2 P_{XGB}\right)$$

where:

1. $P_{RF}$ and $P_{XGB}$: Base learners (Random Forest and XG Boost) calibrated probabilities.
2. $\beta_1$ and $\beta_2$: Weights learned by Logistics Regression meta learner
3. $\sigma(.)$: Sigmoid Function

### F. Proposed Hybrid Disease-Specific Ensemble

Unlike traditional stacked ensembles, which use a global meta-model, this methodology utilizes a Hybrid Disease Aware Ensemble

1. This model assigns disease-specific weights to RF and XGB, depending on their performance for that disease.
2. Predicted probabilities from RF and XGB are combined using these weights to generate a hybrid probability score.
3. To improve probability reliability, RF and XGB were calibrated using isotonic regression before combining [27].

Hybrid ensemble uses calibrated base probabilities with disease specific weights and thresholds, instead of using a global Logistics Regression.

$$P_{hybrid,\,d} = w_d \cdot P_{calibrated\_RF,\,d} + (1 - w_d) \cdot P_{calibrated\_XGB,\,d}$$

where:

1. $P_{hybrid,d}$: Final probability for disease d
2. $w_d$: Disease-specific learned weight
3. $P_{calibrated\_RF,d}$ and $P_{calibrated\_XGB,d}$: Calibrated outputs.

This approach allows the system to favour the algorithm that performs best for a particular disease. This improves recall considerably while accepting a slight reduction in accuracy. This is more desirable because in a real-world scenario false negatives (missed diagnoses) are more expensive than false positives.

### G. Threshold Tuning

Disease-specific thresholds are selected to optimize performance for clinical relevance.

Threshold selection aims to maximize recall while maintaining a minimum specificity of 0.7. This framework helps to balance the sensitivity and specificity. This addresses the clinical need to reduce false positives (misdiagnosing healthy patients) without significantly increasing false negatives (missing actual diagnoses).

Threshold tuning ensures balanced sensitivity/specificity for each disease:

$$\hat{y}_d = \begin{cases} 1 & if\ P_{hybrid,d} \geq \tau_d \\ 0 & otherwise \end{cases}$$

where $\tau_d$ is the optimal threshold chosen for disease d, subject to specificity $\geq 0.7$.

### H. System Workflow

The proposed Hybrid Disease-Aware Ensemble workflow is illustrated in Fig. I.

### I. Evaluation Metrics

1. Accuracy: Measures the proportion of correct predictions a model makes out of the total number of predictions.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

2. Precision: Measures the proportion of positive predictions that were correct.

$$Precision = \frac{TP}{TP + FP}$$

3. Recall (Sensitivity): Measures the proportion of actual positives that were identified correctly.

$$Recall\ (Sensitivity) = \frac{TP}{TP + FN}$$

4. Specificity: Measures the proportion of actual negatives that were identified correctly.

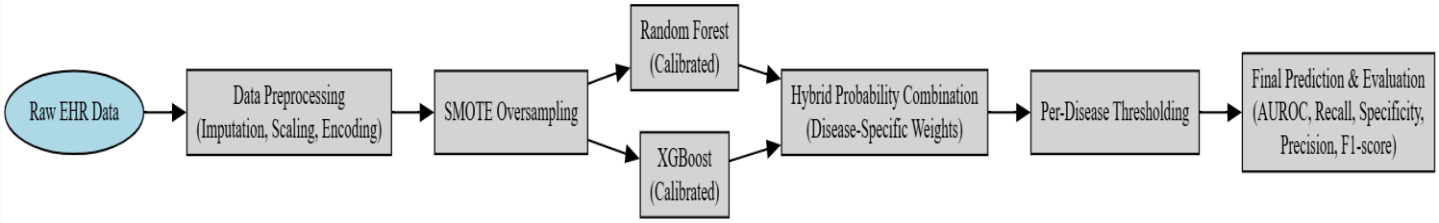
$$Specificity = \frac{TN}{TN + FP}$$

5. F1-Score: The harmonic mean of Precision and Recall.

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Where:
TP = True Positives
TN = True Negatives
FP = False Positives
FN = False Negatives.

FIG. I.: WORKFLOW OF THE HYBRID DISEASE-AWARE ENSEMBLE




## IV. RESULTS AND DISCUSSION

Two ensemble strategies have been evaluated for early disease prediction using Electronic Health Records:

1. Classic Stacked Ensemble Model: This model combines base learners, Random Forest, and XGBoost with a single global meta-learner, Logistic Regression. The implementation is much simpler, and predictions from all base learners are treated uniformly for all diseases; it underperforms for minority diseases or those with distinct features. Table III. shows the evaluation metrics of the model.
2. Proposed Hybrid Disease-Aware Ensemble Model: This model combines RF and XGB by using disease-specific weights calibrated for each disease using the Isotonic Regression Model to produce reliable probability estimates. Table IV. shows the evaluation metrics of the model. Fig. II. represents the ROC curves of each disease, and Fig. III. represents the Confusion Matrix.

The classical stacked ensemble model favoured a high sensitivity at the cost of reduced specificity, leading to a high rate of false positives. For example, in case of Breast Cancer, the classical model achieved perfect sensitivity of 1.000 but with a low specificity of just 0.730, which could result in many healthy patients being incorrectly marked as diseased. In contrast, the proposed ensemble model maintains the perfect sensitivity while improving the specificity considerably to 0.955. This demonstrates the effectiveness of the disease-specific weighting and threshold tuning in reducing false alarms without missing actual diagnoses.

This trend is clear in other cases as well. For general cancer, the hybrid model increased overall accuracy from 0.847 to 0.964 and achieved perfect specificity at 1.000, up from a modest 0.782. For diabetes, the proposed model improved a poor accuracy of 0.785 in the traditional model to 0.805. Although this rebalancing sometimes led to a slight, clinically acceptable decrease in sensitivity, such as from 0.911 to 0.857 for diabetes, the resulting drop in false positives and the increase in overall model reliability, with the F1-score rising from 0.761 to 0.768, show a significant gain in clinical usefulness.

These results demonstrate a contrast between a conventional stacked ensemble and a hybrid stacked ensemble. The traditional stacked ensemble achieves strong predictive performance, but the result is at the cost of an imbalance between sensitivity and specificity. This means many healthy patients have been classified as diseased. This increases the spending on medical resources and slows clinical workflows. The model also displayed a low accuracy.

The Hybrid Disease-Aware Ensemble addresses this limitation by introducing disease-specific weighting and enforcing a minimum specificity threshold. This model improves accuracy of predicting diseases and the balance between sensitivity and specificity. It reduces the number of false positives, which is important for early diagnosis, while also maintaining a low rate of false negatives. Now, unlike the classical ensemble, the hybrid model reduces the risk of classifying too many healthy patients as diseased. While still maintaining a high AUROC and sensitivity. The probabilities have been calibrated using Isotonic Regression, which improves the reliability of the predictions and makes the hybrid model more clinically relevant in real-world scenarios.

### *A. Limitations of the Study*

While this research presents a promising Hybrid Disease – Aware Ensemble Model, it is necessary to address some limitations in our findings to provide ways for future research.

The dataset that has been used for training and evaluation was constructed by combining multiple publicly available disease specific datasets. This approach may introduce a potential heterogeneity. A more robust evaluation would involve using large, standardized EHR datasets such as MIMIC-III or MIMIC-IV as these provide a more consistent representation of clinical data [17][28].

The scope of this research was limited to a few diseases. Only a specific number of diseases have been used in demonstrating the model's performance. Improvements can be made to model's performance for detecting a broader set of diseases.

TABLE III.: EVALUATION METRICS OF THE CLASSICAL STACKED ENSEMBLE MODEL

| Disease | AUROC | Accuracy | Sensitivity/ Recall | Specificity | Precision | F1-Score |
|---|---|---|---|---|---|---|
| Cancer | 0.979 | 0.847 | 0.960 | 0.782 | 0.716 | 0.821 |
| Breast Cancer | 0.992 | 0.824 | 1.000 | 0.730 | 0.662 | 0.797 |
| Lung Cancer | 0.883 | 0.860 | 0.880 | 0.429 | 0.957 | 0.917 |
| Diabetes | 0.880 | 0.785 | 0.911 | 0.971 | 0.654 | 0.761 |
| Heart Disease | 0.880 | 0.817 | 0.903 | 0.724 | 0.778 | 0.836 |
| CKD | 1.000 | 0.930 | 1.000 | 0.828 | 0.894 | 0.944 |

TABLE IV.: EVALUATION METRICS OF THE PROPOSED HYBRID DISEASE-AWARE ENSEMBLE MODEL

| Disease | AUROC | Accuracy | Sensitivity/ Recall | Specificity | Precision | F1-Score |
|---|---|---|---|---|---|---|
| Cancer | 0.975 | 0.964 | 0.900 | 1.000 | 1.000 | 0.947 |
| Breast Cancer | 0.991 | 0.971 | 1.000 | 0.955 | 0.922 | 0.959 |
| Lung Cancer | 0.837 | 0.825 | 0.840 | 0.714 | 0.955 | 0.894 |
| Diabetes | 0.880 | 0.805 | 0.857 | 0.774 | 0.696 | 0.768 |
| Heart Disease | 0.884 | 0.817 | 0.903 | 0.724 | 0.778 | 0.836 |
| CKD | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |

FIG. II.: ROC CURVES PER DISEASE PROPOSED HYBRID DISEASE-AWARE ENSEMBLE MODEL

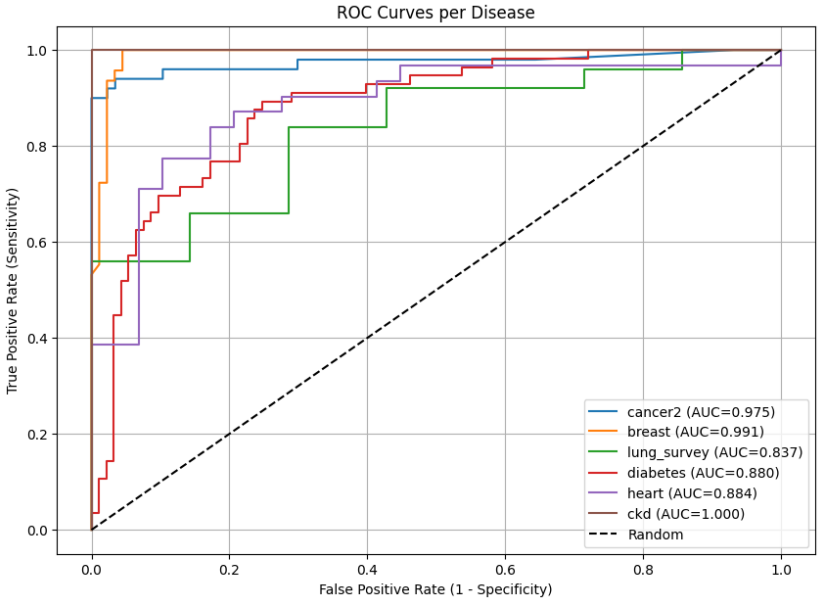
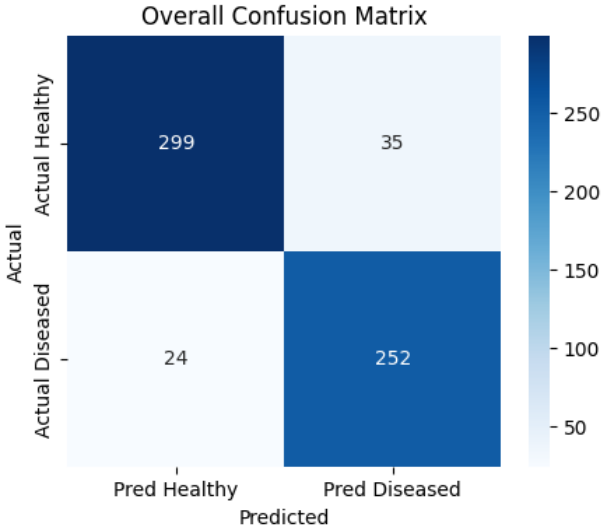



FIG. III.: CONFUSION MATRIX OF THE PROPOSED HYBRID DISEASE-AWARE ENSEMBLE MODEL




## V. CONCLUSION

### A. Summary and Contributions

This paper examined the challenges of different machine learning models used to predict different diseases. To overcome the reliance on multiple models, the research provides a comparative evaluation between two ensemble-based approaches for early disease prediction using electronic health records. The primary contribution of the research is the development of a novel Hybrid Disease Aware Ensemble which uses disease-specific weighing and calibrated thresholds. This model overcomes the limitations of a stacked ensemble model which, despite achieving strong predictive results, suffers from low accuracy and an imbalance between sensitivity and specificity. By reducing false positives, the proposed model ensures fewer healthy patients are misdiagnosed as diseased. This saves valuable medical resources and prevents delays for treating genuinely ill patients.

### B. Challenges and Limitations

Although the results are promising, we acknowledge a few limitations. The model was tested on a synthetic dataset made by aggregating a variety of disease-specific datasets, and it might not capture the full richness of a single, unified clinical setting. Additionally, we restricted ourselves to a few diseases in this study, and for a full range of conditions the performance of the model needs to be studied further.

### C. Future Scope

Future research should investigate using the proposed method on larger, standardized datasets like MIMIC-III or MIMIC-IV [17][28]. Another subsequent step of significance is to expand the framework to cover a wider spectrum of disease classes. A significant opportunity also lies in combining Large Language Models (LLMs) and Natural Language Processing (NLP) to process unstructured data such as

clinical notes, images, and videos [7]. This method could pave the way to facilitate with a robust real-time decision-making in early disease diagnosis.

## VI. REFERENCES

[1] K. P, S. Y, S. V and V. V, "Medical Condition Diagnosis Through the Application of Machine Learning to EHRs," *2024 Second International Conference on Advances in Information Technology (ICAIT)*, Chikkamagaluru, Karnataka, India, 2024, pp. 1-5, doi: https://doi.org/10.1109/icait61638.2024.10690505

[2] B. Kumar Saraswat, A. Saxena and P. C. Vashist, "Opportunities and Challenges for Developing Machine Learning Models with EHR Data," *2023 International Conference on Advanced Computing & Communication Technologies (ICACCTech)*, Banur, India, 2023, pp. 649-656, doi: https://doi.org/10.1109/icacctech61146.2023.00110.

[3] Y. Zhang, J. Liu, H. Liu, Y. Lu, S. Wang, and Y. Zhai, "High Dimensional Missing Data Imputation for Classification Problems: A Hybrid Model based on K-Nearest Neighbor and Genetic Algorithm," *2022 International Symposium on Advances in Informatics, Electronics and Education (ISAIEE)*, pp. 572–578, Dec. 2022, doi: https://doi.org/10.1109/isaiee57420.2022.00121.

[4] L. Xu and A. Qiu, "Multiple Imputation by Chained Equations for Missing Data in UK Biobank," *2022 6th Annual International Conference on Data Science and Business Analytics (ICDSBA)*, Changsha, China, 2022, pp. 72-82, doi: https://ieeexplore.ieee.org/document/10129436

[5] K. M. Sujon, R. Hassan and N. Jahan, "Synthetic Minority Over-sampling Technique for Student Performance Prediction: A Comparative Analysis of Ensemble and Linear Models," *2024 27th International Conference on Computer and Information Technology (ICCIT)*, Cox's Bazar, Bangladesh, 2024, pp. 2231-2236, doi: https://doi.org/10.1109/iccit64611.2024.11022420.

[6] C. A. Stevens *et al.*, "Ensemble machine learning methods in screening electronic health records: A scoping review," *DIGITAL HEALTH*, vol. 9, Jan. 2023, doi: https://doi.org/10.1177/20552076231173225.

[7] A. Sarker *et al.*, "Natural Language Processing for Digital Health in the Era of Large Language Models," *Yearbook of Medical Informatics*, vol. 33, no. 01, pp. 229–240, Aug. 2024, doi: https://doi.org/10.1055/s-0044-1800750.

[8] S. Ara, A. Das and A. Dey, "Malignant and Benign Breast Cancer Classification using Machine Learning Algorithms," *2021 International Conference on Artificial Intelligence (ICAI)*, Islamabad, Pakistan,2021,pp.97-101, doi: https://doi.org/10.1109/ICAI52203.2021.9445249

[9] A. Rovshenov and S. Peker, "Performance Comparison of Different Machine Learning Techniques for Early Prediction of Breast Cancer using Wisconsin Breast Cancer Dataset," *2022 3rd International Informatics and Software Engineering Conference (IISEC)*, Ankara, Turkey,2022,pp.1-6, doi: https://doi.org/10.1109/iisec56263.2022.9998248.

[10] P. S, V. B, L. Krishnasamy, T. P, P. R. M and S. S, "Lung Cancer Prediction using Machine Learning," *2025 3rd International Conference on Communication, Security, and Artificial Intelligence (ICCSAI)*, Greater Noida, India, 2025, pp. 1604-1608, doi: https://doi.org/10.1109/iccsai64074.2025.11063814.

[11] P. Verma and A. Khatoon, "Data Mining Applications in Healthcare: A Comparative Analysis of Classification Techniques for Diabetes Diagnosis Using the PIMA Indian Diabetes Dataset," *2024 4th International Conference on Innovative Practices in Technology and Management (ICIPTM)*, pp. 1–5, Feb. 2024, doi: https://doi.org/10.1109/iciptm59628.2024.10563296.

[12] M. A. Bouqentar *et al.*, "Early Heart Disease Prediction using Feature Engineering and Machine Learning Algorithms," *Heliyon*, vol. 10, no. 19, pp. e38731–e38731, Oct. 2024, doi: https://doi.org/10.1016/j.heliyon.2024.e38731.

[13] Anurag, N. Vyas, V. Sharma and D. Balla, "Chronic Kidney Disease Prediction Using Robust Approach in Machine Learning," *2023 3rd International Conference on Innovative Sustainable Computational Technologies (CISCT)*, Dehradun, India, 2023, pp. 1-5, doi: https://doi.org/10.1109/cisct57197.2023.10351277.

[14] N. H. Alhumaidi, D. Dermawan, H. F. Kamaruzaman, and N. Alotaiq, "The Use of Machine Learning for Analyzing Real-World Data in Disease Prediction and Management: Systematic Review," JMIR Medical Informatics, vol. 13, p. e68898, Jun. 2025, doi: https://doi.org/10.2196/68898.

[15] D. Pei, C. Zhang, Y. Quan, and Q. Guo, "Identification of Potential Type II Diabetes in a Chinese Population with a Sensitive Decision Tree Approach," *Journal of Diabetes Research*, vol. 2019, pp. 1–7, Jan. 2019, doi: https://doi.org/10.1155/2019/4248218.

[16] Abubakar, B. B. (2025). The Role of Domain Knowledge in Feature Selection for Machine Learning. Journal of Institutional Research, Big Data Analytics and Innovation, 1(3), 180–187. https://doi.org/10.5281/zenodo.16615429

[17] Johnson, A., Bulgarelli, L., Pollard, T., Horng, S., Celi, L. A., & Mark, R. (2023). MIMIC-IV Clinical Database Demo (version 2.2). *PhysioNet*. RRID:SCR_007345. https://doi.org/10.13026/dp1f-ex47.

[18] A. an and R. S. Chhillar, "Disease Predictive Models for Healthcare by using Data Mining Techniques: State of the Art," International Journal of Engineering Trends and Technology, vol. 68, no. 10, pp. 52–57, Oct. 2020, doi: https://doi.org/10.14445/22315381/ijett-v68i10p209.

[19] M. James, "Comparative Analysis of Machine Learning Models for Early Detection of Heart Disease from EHRs," *ResearchGate*, Jun. 24, 2024. https://www.researchgate.net/publication/395014607

[20] H.-H. Rau *et al.*, "Development of a web-based liver cancer prediction model for type II diabetes patients by using an artificial neural network," *Computer Methods and Programs in Biomedicine*, vol. 125, pp. 58–65, Mar. 2016, doi: https://doi.org/10.1016/j.cmpb.2015.11.009.

[21] R. Sundar *et al.*, "Machine-learning model derived gene signature predictive of paclitaxel survival benefit in gastric cancer: results from the randomised phase III SAMIT trial," *Gut*, vol. 71, no. 4, pp. 676–685, May 2021, doi: https://doi.org/10.1136/gutjnl-2021-324060.

[22] S. M. Ganie, P. K. D. Pramanik, and Z. Zhao, "Ensemble learning with explainable AI for improved heart disease prediction based on multiple datasets," Scientific Reports, vol. 15, no. 1, Apr. 2025, doi: https://doi.org/10.1038/s41598-025-97547-6.

[23] C. Choudhary, L. S. Nagra, P. Das, J. Singh and S. S. Jamwal, "Optimized Ensemble Machine Learning Model for Chronic Kidney Disease Prediction," *2023 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS)*, Greater Noida, India,2023,pp.292-297, doi: https://doi.org/10.1109/icccis60361.2023.10425073.

[24] R. Rani, "Optimized Heart Failure Prediction using Support Vector Machine Algorithms," *2024 5th International Conference on Smart Electronics and Communication (ICOSEC)*, Trichy, India, 2024, pp. 1265-1268, doi: https://doi.org/10.1109/icosec61587.2024.10722131.

[25] S. Rallapalli and T. Suryakanthi, "Predicting the risk of diabetes in big data electronic health Records by using scalable random forest classification algorithm," *2016 International Conference on Advances in Computing and Communication Engineering (ICACCE)*, Nov. 2016, doi: https://doi.org/10.1109/icacce.2016.8073762.

[26] L. Anandhan, N. A., S. A. P., R. V., and S. Sanjay, "Personalized Drug Information and Recommendation System Using Gradient Boosting Algorithm (GBM)," *2024 13th International Conference on System Modeling & Advancement in Research Trends (SMART)*, pp. 923–926, Dec. 2024, doi: https://doi.org/10.1109/smart63812.2024.10882477.

[27] M. P. Naeini and G. F. Cooper, "Binary Classifier Calibration Using an Ensemble of Near Isotonic Regression Models," *2016 IEEE 16th International Conference on Data Mining (ICDM)*, Barcelona, Spain, 2016, pp. 360-369, doi: https://doi.org/10.1109/icdm.2016.0047.

[28] A. Johnson *et al.*, "OPEN SUBJECT CATEGORIES Background & Summary," *MIMIC-III, a Freely Accessible Critical Care Database*, 2016, doi: https://doi.org/10.1038/sdata.2016.35.