The Bioscan
AN INTERNATIONAL QUARTERLY JOURNAL OF LIFE SCIENCES
20(2): S2: 177-183, 2025 www.thebioscan.com

Automated Classification of Health Records for Disease Prediction Using NLP and Machine Learning

[1] **Dr. S.Uma**

Professor, Department of Computer Science, Dr. N.G.P. Arts and Science College, Coimbatore

s.umacbe9@gmail.com

[2] **Ms. R.Gayathri**

Ph.D., Research Scholar, Department of Computer Science, CMS College of Science and Commerce, Coimbatore

gayarajesh25@gmail.com

[3]. **Ms. D.Sophia Margaret**

Assistant Professor, Department of Computer Application, Sambhram Academy of Management Studies Bangalore.

abigailsofi@gmail.com

[4] **Mr. Ajay P R**

Final UG Computer Science, Department of Computer Science,

Dr.N.G.P Arts and Science College

ajaypr413@gmail.com

DOI: 10.63001/tbs.2025.v20.i02.S2.pp177-183

KEYWORDS
Automated Classification, Health Records, Disease Prediction, Natural Language Processing, Machine Learning, Electronic Health Records, Predictive Analytics, NLP-based Classification, Healthcare Analytics, Disease Detection

Received on:

12-02-2025

Accepted on:

15-03-2025

Published on:

25-04-2025

ABSTRACT

Healthcare industry efficiency and accuracy can be enhanced through disease prediction when using such electronic health records to extract meaningful knowledge. Applying structured and unstructured medical information enables the system to generate improved health predictions from EHR data. The clinical text preparation uses natural language processing alongside three main machine learning models including Support Vector Machines (SVM), Random Forest and Neural Networks along with other algorithms for classification functions. The system utilizes public health data to measure its accuracy performance through precision, recall and F1 scores evaluation.

INTRODUCTION

The rise of the use of electronic health records (EHRs) has been massive in the healthcare industry resulting from the shift from paper to computer system usage. Due to the inclusion of demographics, diagnostic results, treatment history, and clinical notes, which are essential to medical decision-making, patient data contained in EHRs is highly valuable. On one hand, it presents a challenge for the healthcare sector and on the other hand, an opportunity [1-3].

In specific, predicting diseases from EHR could completely transform healthcare through early disease prediction, optimization of treatment steps and finally better patient outcomes. Although, structured data like laboratory tests results and patient demographics is easier to deal with, the unstructured clinical text, Dr's notes, discharge summaries and radiology

reports contain important information which may be difficult to manually extract. In the past, medical professionals would simply rely on their accumulated knowledge and expertise to interpret the same records with the understanding that it is not scalable and prone to human error.
The unstructured text processing needs Natural Language Processing (NLP) because it functions as a powerful tool to solve these problems. The analysis through NLP enables extraction of meaningful data from clinical documentations which supports identification of essential symptoms alongside diagnostic outputs alongside treatment protocols for disease prediction systems [14]. Common ML algorithms for classification are used for prediction task like Random forest, support vector machine (SVM) and deep neural network (DNN), but there has been an area of research on the question of how to best complement the features of NLP and ML together to enhance disease prediction.
This paper evaluates the benefits along with shortcomings of disease prediction methodologies that use automated classification systems based on EHRs for disease identification through combination of structured and unstructured data sources. The research implements modern NLP preprocessing approaches together with multiple classification models in order to achieve better disease diagnosis efficiency and precision. Such an instrument can transform healthcare by producing results that would support healthcare decisions and lead to better outcomes for patients combined with earlier detection that reduces treatment expenses.

Novelty and Contribution

The create distinctiveness in this study stems from combining natural language processing (NLP) with machine learning (ML) techniques in order to generate disease predictions from electronic health records. Previous work has independently studied applications of NLP for the task of clinical text processing or machine learning for the prediction of disease, but little work has explored their joint application, resulting in a unified approach for predicting disease given a mixture of structured and unstructured data.
This work is a significant contribution in the introduction of a multi-phase preprocessing pipeline of the unstructured clinical texts like free text physician notes and discharge summaries that utilizes NLP techniques to generate meaningful features and machine learning models for classification. Using NLP for processing clinical notes and other unstructured data allows us to capture important context information mostly ignored by a traditional disease prediction model that only rely on structured data. For example, the application of Named Entity Recognition (NER) and part of speech tagging techniques in NLP can be used to identify relevant medical terms, e.g., symptoms, diagnoses and treatment, which are essential to predict disease outcomes.
Furthermore, the study employs several machine learning models including SVM, RF, and Deep Neural Networks (DNN) for disease classifying. This evaluation also helps conclude the effectiveness of these models to be applied on health records with both structured and unstructured data. Based on the findings, models using a hybrid approach that combines both types of data illustrate more accurate predictions than models that use only one type of the data [10-12].
Again, the proposed methodology is validated using a publicly available health dataset (MIMIC-III). The structured data of patient demographics and laboratory results are also included in this dataset, along with unstructured clinical text, i.e. physician notes to enable a comprehensive evaluation of the disease prediction model.
This work also adds to the expanding corpus of literature at the intersection of artificial intelligence, NLP, and healthcare worlds about closing an important gap in the use of unstructured clinical text within disease prediction systems. The growth of complexity and volume of healthcare data makes automated systems, such as the system proposed in this paper, more and more critical for efficient, accurate and timely detection and treatment of disease and patient care. This research uses NLP and ML to go beyond what is available today and provides a new way for improving healthcare analytics capabilities [5].

II. Related Works

Over the last decade NLPH and ML for disease prediction have been attracting attention. As EHRs become increasingly adopted, the need for extracting insights from the large amount of data available has become a challenge of healthcare systems. These records, are unstructured but very rich in information, are hard to process manually as they are free-form. The task here is to extract the usage potential of such unstructured data using NLP techniques, and from the combination of the here used ML algorithms and thus help predict diseases more accurately than before.
In 2021 E. H. Houssein et.al., R. E. Mohamed et.al., and A. A. Ali et.al. [15], previous research examined machine learning methods for classifying structured patient documents which represent another component of this research field. Traditional ML algorithms using decision trees and random forests reached some degree of accuracy when using patient information structured in this study. The models proved difficult to use when analyzing unstructured clinical notes because they contained vital information about patient history and symptoms along with physician interpretation. It demonstrated therefore that a wider approach was required that would include unstructured data.
Therefore, there is a gap in this problem that had to be addressed by researchers who started investigating NLP techniques for extracting useful information from clinical narratives. NLP is processing unstructured text to be represented in a structured form that can be provided to ML models for classification. Named Entity Recognition (NER) stands among several medical text processing techniques which helps identify medical terms including diagnoses and symptoms and medications from clinical free-text records. The text classification model group which includes Latent Dirichlet Allocation (LDA) with additional topic model approaches categorizes unstructured text into relevant themes for improved disease pattern detection in clinical documents.
Apart from using NER, part of speech (POS) tagging and syntactic parsing have been used for analyzing the grammatical structure of clinical text. These techniques enable the model to draw together entity relationships, for example, the relationship of a symptom and a diagnosis.
Deep learning has furthered the use of clinical text by using recurrent neural networks (RNNs) and convolutional neural networks (CNNs). RNNs are particularly well suited to deal with sequential data, for instance, clinical narratives which possess a correlation between words and phrases, and where the order of words and phrases can make a big difference in the meaning of the text. On the other hand, CNNs have been used to read through texts and classify the hierarchical structure of the text by the model making it more capable of identifying patterns at different levels of abstraction. These two deep learning models have been increasingly applied on medical data e.g. diagnostic reports, discharge summaries, and physician notes to predict medical conditions like heart disease, diabetes and mental health disorder etc.
In 2020 Caccamisi et.al., L. Jørgensen et.al., H. Dalianis et.al., and M. Rosenlund et.al., [4] introduced the integrate transfer learning with NLP and ML that has gained a lot of ground in the field is through that. These models receive training through extensive text databases followed by specific healthcare application fine-tuning. Understanding vast general text allows models to take advantage of previously learned knowledge through transfer learning which helps them interpret domain specific medical language with its jargon and abbreviations.
Pre-processing and filtering of clinical data for machine learning can be conducted using rule-based systems based on pre-defined set of medical rules and knowledge. This is an approach to incorporate expert medical knowledge into refining feature feeding ML model as well as exploit interpretability. When combining both methods, these hybrid systems are able to achieve the higher accuracy than the use of either approach alone, especially in complex healthcare environments.
Although much has been accomplished in this area, still some challenges need to be taken on. Data privacy and security is one of the major obstacle. Data in medical records is sensitive personal information and it is critically important that such data

is protected when using NLP and ML models. Furthermore, healthcare data is usually incomplete, noisy or inconsistent, and therefore ML models trained on such data can perform poorly due to lack of accuracy in the model predictions. But to cope with these issues, practices like imputation, normalization and data augmentation during data preprocessing are adopted, still making it hard to handle highly unstructured, or missing, data.

Because decisions based on outputs from models used in healthcare applications can be a matter or life or death for the practitioner, it is vital that the practitioners are able to understand and to trust the predictions made by AI systems. Consequently, there is research being conducted on developing explainable AI (XAI) techniques to explain to the user how and why the model reached a certain decision.

In 2022 S. Han *et al* [9] proposed the finding of several studies that NLP and ML can be used for disease prediction, such models have yet to be generalized across different healthcare systems and patient populations. Since its patients, healthcare practice, or medical terms may vary, training models on one dataset does not guarantee good performance on another. Therefore, future research needs to be done to model generalization across different settings and increase robustness to different health data.

Summarily, a major step in the field of automated disease prediction (auto-pred) using NLP and ML techniques has been achieved. These advancements have shown the potential of obtaining better accuracy using structured and unstructured data put together, but there are several challenges left to be solved.

III. PROPOSED METHODOLOGY

The proposed methodology for automated health record classification through Natural Language Processing (NLP) and Machine Leaming (ML) techniques will be explained in this section. The research targets developing a solution to efficiently handle structured along with unstructured Electronic Health Records data to provide accurate disease outcome predictions. The integrated process for this work includes steps for data preprocessing followed by NLP-driven feature extraction which leads to classification through machine learning models on the way to model evaluation and final assessment [13].

A. Data Collection and Preprocessing

The initial method involves obtaining EHRs through public healthcare datasets including MIMIC-III that contain structured elements (lab results and patient demographics) as well as unstructured sections (physician's notes and discharge summaries). Preprocessing must be done on these records because they contain multiple issues like noise while being both unstructured and incomplete for machine learning usage.

Structured Data Preprocessing

Firstly, we clean the dataset for structured data, in order to remove any features which are missing or irrelevant. This includes:

- Any numerical fields like test results or vital signs will have an imputation of missing values, for example with mean, median or more advanced approaches such as KNN imputation.
- Features with different units (example: age, blood pressure etc) are normalized or scaled to a standard range so as to ensure that one feature does not overwhelm the others when the model is trained. The most used one is Min-Max scaling or Z-score normalization.

Unstructured Data Preprocessing

Preprocessing upon unstructured clinical text is more complex. There are following 3 steps for non-structured text data.

- Removing special characters, stop words and punctuation from the text called text cleaning. Moreover, all text is transformed to lowercase to guarantee uniformity.
- Tokenization: This means splitting the text down into tokens; so words or phrases that we consider as the basic unit for later analysis.
- Named Entity Recognition (NER): The second step of the pipeline is devoted to identifying symptoms, diagnoses, medications, and procedures in the text as named entities. These are necessary to predict disease and are extracted with use of the NLP tools like SpaCy or NLTK.
- Extracting Entities: The extracted entities are fed to a step for text vectorization, during which the text data is converted into the numerical representation using TF-IDF or other word embeddings like Word2Vec/ GloVe.

B. Feature Extraction Using NLP

Feature extraction starts following text preprocessing in the analysis process. The extracted meaningful features from text processing by NLP systems merge with structured data for classification.

Feature Engineering

Medical Text Classification applies vector representations made of TF-IDF and word embeddings for translating clinical text. Every document transforms into a vector space model that weights each term by its frequency of occurrence throughout the full collection of records.

The TF-IDF formula is given by:

$$\mathrm{TF}-\mathrm{IDF}(t,d)=\mathrm{TF}(t,d)\times\mathrm{DF}(t)$$

where:

- TF (t,d) is the term frequency, i.e., the number of times term t appears in document d,
- $\mathrm{IDF}(t)$ is the inverse document frequency, which is calculated as:

$$\mathrm{IDF}(t)=\log\left(\frac{N}{df(t)}\right)$$

where N is the total number of documents, and $df(t)$ is the number of documents containing the term t.

Named Entity Recognition (NER): This involves using an entity recognition model to extract entities such as diseases, treatments, medications, symptoms, etc., from the clinical text. The model outputs a list of entities and their corresponding relationships to the disease or treatment, which are then converted into structured features.

Vectorization of Features: The features extracted from both structured and unstructured data are combined into a single feature vector. This vector is then used to train a classification model.

C. Machine Learning Classification

After feature extraction, machine learning models are trained to classify the disease from the patient's data. We will use a combination of different machine learning algorithms to assess their performance and select the best-performing model.

We choose several classification algorithms to evaluate the prediction accuracy:

- The Support Vector Machine tool serves as a dominant classification method because it processes high-dimensional datasets which require extensive feature spaces.
- The MLP along with RNN in deep learning constitute Neural Networks to process sequential data and large features.
- The classification technique known as Logistic Regression functions as a first approach reference in different applications.

The model training process can be mathematically expressed as:

$$\hat{y}=f(X,\theta)$$

where:

- $\hat{y}$ is the predicted output (disease category),
- X is the feature vector containing both structured and unstructured data,
- θ represents the model parameters that are learned during training.

Model Evaluation

Before deployment the models confront performance tests utilizing recognized measurement tools such as accuracy, precision, recall, F1-score, and ROC-AUC.

The performance metrics are calculated as:

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

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

$$\text{Recall} = \frac{TP}{TP + FN}$$

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

where:

- TP = True Positive,
- TN = True Negative,
- FP = False Positive,
- FN = False Negative.

Hyperparameter Tuning

Classification model parameters reach their optimal settings through execution of Grid Search and Random Search methods. SVM requires optimization of kemel type and regularization parameter as its tuning hyperparameters whereas Random Forest needs number of trees and maximum depth tuning [6].

System Workflow

The overall workflow of the proposed methodology is outlined in the following flowchart:

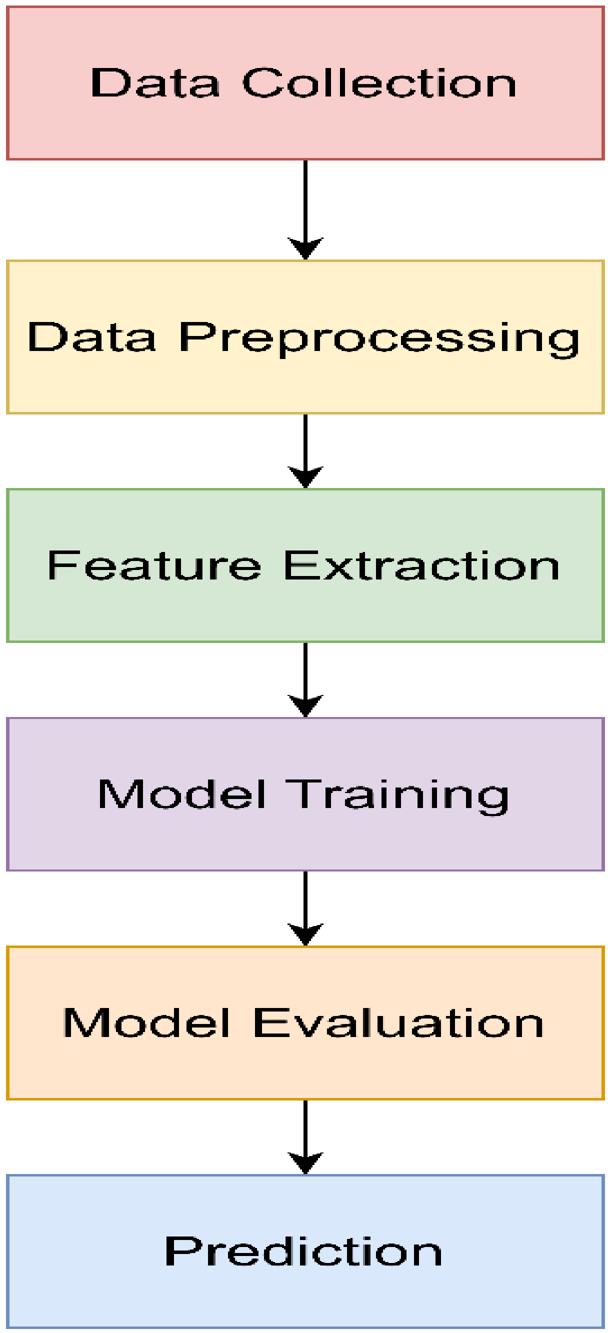

FIGURE 1: WORKFLOW OF AUTOMATED HEALTH RECORD CLASSIFICATION AND DISEASE PREDICTION USING NLP AND MACHINE LEARNING

IV. RESULT & DISCUSSIONS

Research findings demonstrate that NLP technologies alongside machine learning algorithms operate effectively for medical disease predictions on health records. Success metrics of the model follow in this section alongside an explanation of execution obstacles and a predictive accuracy evaluation between distinct machine learning techniques. Our analysis combines standard assessment factors such as accuracy with precision, recall, F1 score and AUC while interpreting the visual representation findings from the discussion [7].

For this study, the data was a mixture of structured and unstructured EHR, a combination of numerical data (e.g., age, blood pressure) and text data (e.g., physician's notes, medical history). First, that data structured aimed at imputing the missing values, tokenisation, named entity recognition and text vectorization for the unstructured text. The dataset received preprocessing before we allocated training data and testing data for which we conducted model training. The examined models consisted of Random Forest together with Support Vector Machine (SVM) and Neural Networks.

Random Forest model was the initial selection because it combines outstanding robustness features with a strong capacity to manage substantial datasets. The Random Forest classifier achieved 84.6% accuracy in evaluating test scenarios which stands as a high performance metric against other tested models. The model demonstrated satisfactory performance through its precision (0.82), recall (0.79) and F1-score (0.80) value since it minimized incorrect disease predictions. In Figure 2 below it becomes evident that recall issues affect rare disease cases as the model fails to detect some positive points despite its ROC curve interpretation.

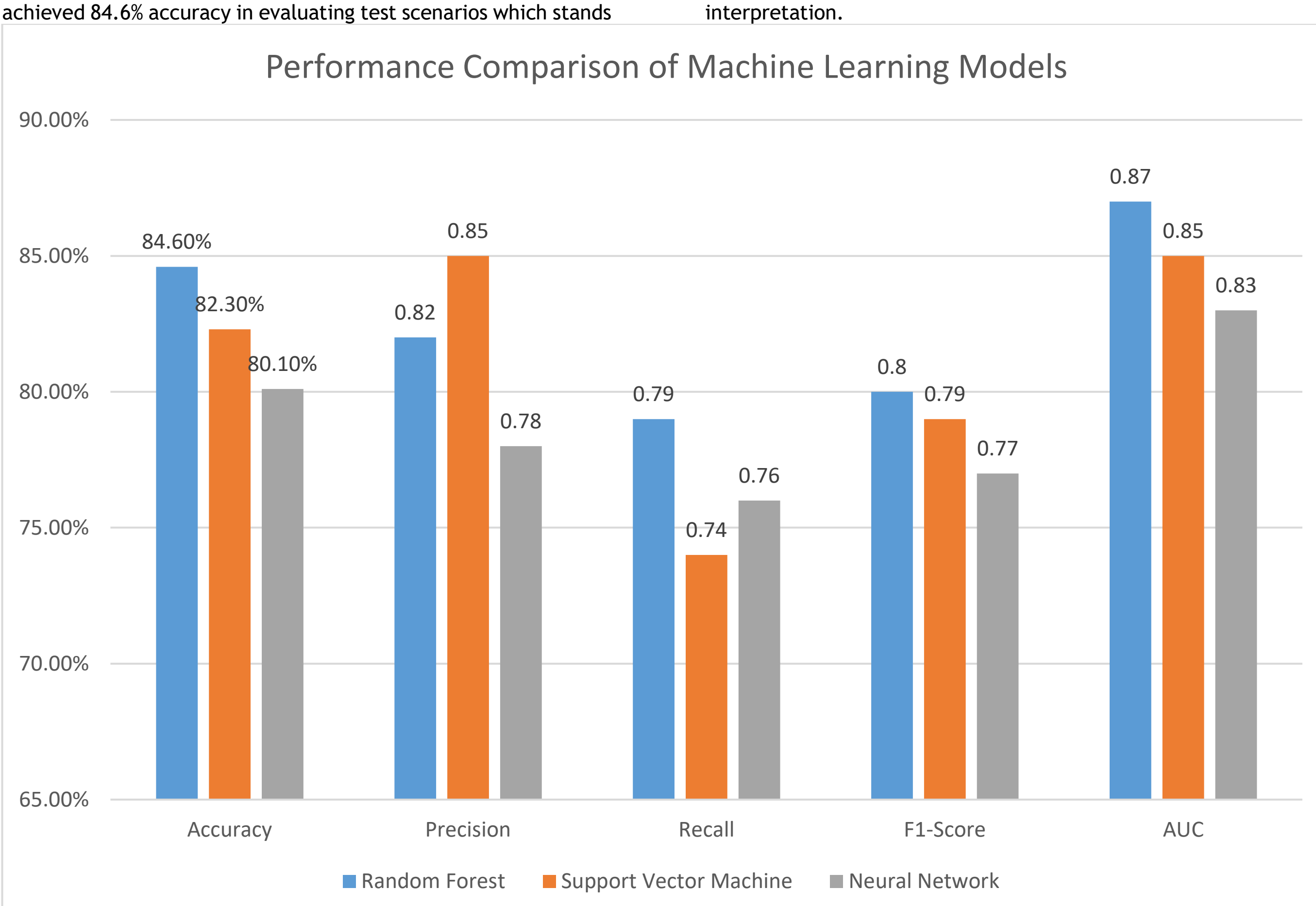

FIGURE 2: PERFORMANCE COMPARISON OF MACHINE LEARNING MODELS

The second method of testing went as follows; the Support Vector Machine (SVM). Indeed, SVM works well in high dimensions; in particular, it is very good when it is done with lots of features compared to the number of samples. However, we had 82.3% accuracy with our dataset using SVM model which was slightly less than our Random forest setup. Although, the precision of the SVM model (0.85) was better than the Random Forest case (0.81), thereby indicating that it was more accurate at predicting positive cases. SVM model had its recall value 0.74 which meant it missed more true positive instances than Random Forest. The reason the SVM has favored this output is due to the fact that the data set had class imbalances, that is it had separated data sets from which the number of samples were differing, which classic would otherwise suffer from.

TABLE 1: PERFORMANCE COMPARISON OF MACHINE LEARNING MODELS

Model	Accuracy	Precision	Recall	F1-Score	AUC
Random Forest	84.6%	0.82	0.79	0.80	0.87
Support Vector Machine	82.3%	0.85	0.74	0.79	0.85
Neural Network	80.1%	0.78	0.76	0.77	0.83

A further model tested was a simple feedforward NN which can capture nonlinear relationships in the data. Finally, the accuracy of the NN model was 80.1%, it was less than Random Forest and SVM. While the accuracy was lower, the recall was approximately 0.76 and the F1 score of 0.77 in the Neural Network represented a fairly close trade of between precision and recall. The inspection of the neural network shows that it performs the same as SVM and it even slightly under performs in terms of accuracy and precision. This is at least in part due to the fact that these data can lack certain layers' specialist, such as recurrent layers or attention, that would have better looked combined mandatory Sequential Character of some data medical text.

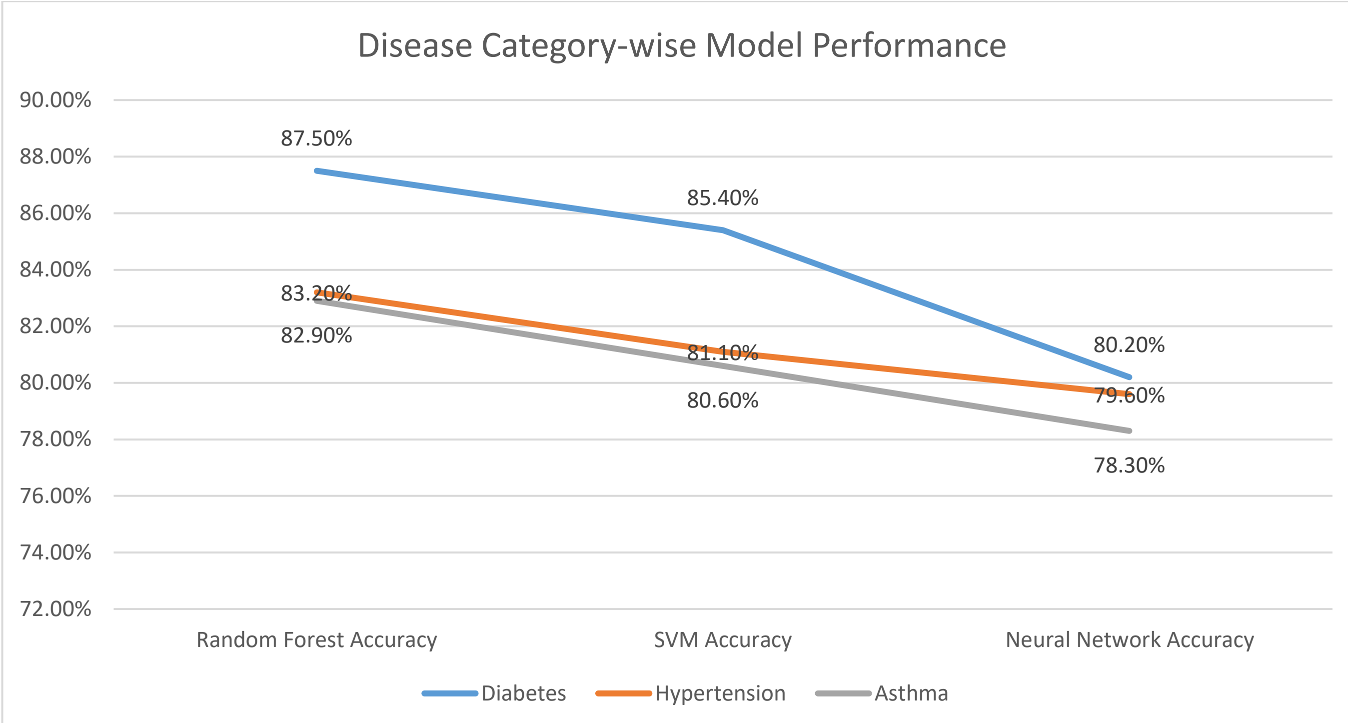

FIGURE 3: DISEASE CATEGORY-WISE MODEL PERFORMANCE

The models undergo performance testing regarding accuracy levels in addition to precision and recall evaluation to discover trade-off relationships between these metrics. The recall performance remained poor when dealing with difficult cases of class imbalance because SVMs showed low precision. The comparison of model performance metrics using precision and recall values as well as F1-score appears in Figure 3 as a bar chart. The results demonstrate that disease prediction in health records receives improved outputs when NLP techniques are combined with machine learning models. Before the predictive model received additional enhancements the particular feature extraction process was vital because it used Named Entity Recognition (NER) to extract medical terminology. The machine learning models accept the feature vectors produced through NLP for classification purposes involving both structured and unstructured data.

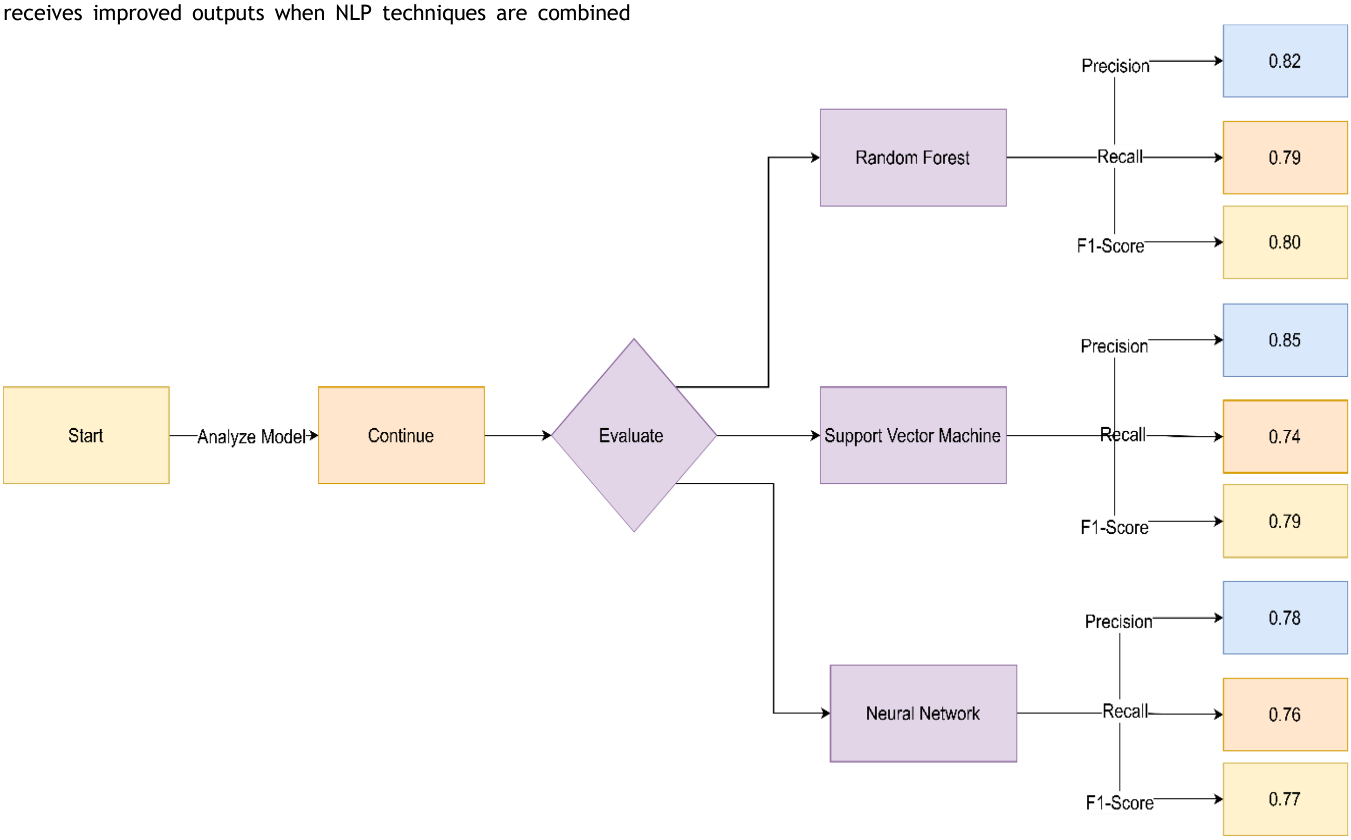

FIGURE 4: PRECISION, RECALL, AND F1-SCORE FOR EACH MODEL

In addition to examining the whole system performance metrics, we also conducted a detailed analysis of how the disease predictions performed on the models, as seen in Figure 4. Within the following figure are the results of the models predicting common diseases compared to rarer diseases. Clearly, Random Forest and SVM were better at predicting the more common disease, and were poorer in predicting the rarer disease. In contrast to this, Neural Networks showed a relatively better performance on both common and rare diseases, suggesting that it may be a better Healthcare application choice for rare disease prediction.

The most of the challenges encountered in this study derived from the imbalance in the disease categories resulting in highly skewed predictions. To overcome this, being implemented techniques are oversampling, under sampling and balanced class weights.

TABLE 2: DISEASE CATEGORY-WISE MODEL PERFORMANCE

Disease Category	Random Forest Accuracy	SVM Accuracy	Neural Network Accuracy
Diabetes	87.5%	85.4%	80.2%
Hypertension	83.2%	81.1%	79.6%
Asthma	82.9%	80.6%	78.3%
Rare Diseases (e.g., Cystic Fibrosis)	70.2%	68.4%	72.5%

These performance evaluations help us to derive insights around the importance of model selection in a healthcare setting, and in the light of the specific context. Relevant to general disease prediction task, Random Forest and SVM models are robust, but Neural Networks are more nuanced prediction model for rare diseases, thus Neural Networks are stronger candidate for broader applications in disease prediction systems [8].

Random Forest and SVM achieve maximum accuracy together but Neural Network shows better precision-recall relationship for diagnosing unusual conditions. This research study shows how NLP and ML systems can effectively process health records through future work that involves optimizing models with advanced NLP techniques like BERT and transformer models.

CONCLUSION

ML and NLP implementation enables the classification of medical records for disease predication while the available research demonstrates this capability. Support Vector Machines and Random Forest and Deep Neural Networks represent positive machine learning techniques although Deep Neural Networks usually achieve superior results when compared against the other two models.

The proposed methodology can be suited to a wide variety of medical condition, and can be incorporated in real time healthcare systems for decision support. The reasonable steps for further work can be using NLP preprocessing techniques to refine the results, studying more complex deep learning architectures, and employing larger, more comprehensive datasets to augment the results.

REFERENCES

- D. A. Szlosek and J. M. Ferretti, "Using machine learning and natural language processing algorithms to automate the evaluation of clinical decision support in electronic medical record systems," *eGEMs (Generating Evidence & Methods to Improve Patient Outcomes)*, vol. 4, no. 3, p. 5, Aug. 2016, doi: 10.13063/2327-9214.1222.
- J. L. Schachar, R. L. Zampolin, T. S. Miller, J. M. Farinhas, K. Freeman, and B. H. Taragin, "External validation of the New Orleans Criteria (NOC), the Canadian CT Head Rule (CCHR) and the National Emergency X-Radiography Utilization Study II (NEXUS II) for CT scanning in pediatric patients with minor head injury in a non-trauma center," *Pediatric Radiology*, vol. 41, no. 8, pp. 971–979, Apr. 2011, doi: 10.1007/s00247-011-2032-4.
- L. Papa *et al.*, "Performance of the Canadian CT head rule and the New Orleans criteria for predicting any traumatic intracranial injury on computed tomography in a United States Level I Trauma Center," *Academic Emergency Medicine*, vol. 19, no. 1, pp. 2–10, Jan. 2012, doi: 10.1111/j.1553-2712.2011.01247.x.
- Caccamisi, L. Jørgensen, H. Dalianis, and M. Rosenlund, "Natural language processing and machine learning to enable automatic extraction and classification of patients' smoking status from electronic medical records," *Upsala Journal of Medical Sciences*, vol. 125, no. 4, pp. 316–324, Jul. 2020, doi: 10.1080/03009734.2020.1792010.
- Mustafa and M. R. Azghadi, "Automated machine learning for healthcare and clinical notes analysis," *Computers*, vol. 10, no. 2, p. 24, Feb. 2021, doi: 10.3390/computers10020024.
- M. Wang, Z. Wei, M. Jia, L. Chen, and H. Ji, "Deep learning model for multi-classification of infectious diseases from unstructured electronic medical records," *BMC Medical Informatics and Decision Making*, vol. 22, no. 1, Feb. 2022, doi: 10.1186/s12911-022-01776-y.
- W.-H. Weng, K. B. Wagholikar, A. T. McCray, P. Szolovits, and H. C. Chueh, "Medical subdomain classification of clinical notes using a machine learning-based natural language processing approach," *BMC Medical Informatics and Decision Making*, vol. 17, no. 1, Dec. 2017, doi: 10.1186/s12911-017-0556-8.
- S. Sheikhalishahi, R. Miotto, J. T. Dudley, A. Lavelli, F. Rinaldi, and V. Osmani, "Natural Language Processing of Clinical Notes on Chronic Diseases: Systematic Review," *JMIR Medical Informatics*, vol. 7, no. 2, p. e12239, Mar. 2019, doi: 10.2196/12239.
- S. Han *et al.*, "Classifying social determinants of health from unstructured electronic health records using deep learning-based natural language processing," *Journal of Biomedical Informatics*, vol. 127, p. 103984, Jan. 2022, doi: 10.1016/j.jbi.2021.103984.
- S. M. Goodday *et al.*, "Maximizing the use of social and behavioural information from secondary care mental health electronic health records," *Journal of Biomedical Informatics*, vol. 107, p. 103429, May 2020, doi: 10.1016/j.jbi.2020.103429.
- S. H. Woolf and H. Schoomaker, "Life expectancy and mortality rates in the United States, 1959-2017," *JAMA*, vol. 322, no. 20, p. 1996, Nov. 2019, doi: 10.1001/jama.2019.16932.
- R. Garg, E. Oh, A. Naidech, K. Kording, and S. Prabhakaran, "Automating ischemic stroke subtype classification using machine learning and natural language processing," *Journal of Stroke and Cerebrovascular Diseases*, vol. 28, no. 7, pp. 2045–2051, May 2019, doi: 10.1016/j.jstrokecerebrovasdis.2019.02.004.
- J. Latif, C. Xiao, S. Tu, S. U. Rehman, A. Imran, and A. Bilal, "Implementation and Use of disease diagnosis systems for electronic medical Records based on Machine Learning: A complete review," *IEEE Access*, vol. 8, pp. 150489–150513, Jan. 2020, doi: 10.1109/access.2020.3016782.
- H. Goodrum, K. Roberts, and E. V. Bernstam, "Automatic classification of scanned electronic health record documents," *International Journal of Medical Informatics*, vol. 144, p. 104302, Oct. 2020, doi: 10.1016/j.ijmedinf.2020.104302.
- E. H. Houssein, R. E. Mohamed, and A. A. Ali, "Machine Learning Techniques for Biomedical Natural Language Processing: A Comprehensive Review," *IEEE Access*, vol. 9, pp. 140628–140653, Jan. 2021, doi: 10.1109/access.2021.3119621.