Decoding Prognosis: A Symptom-Driven Approach to Predictive Healthcare Using Random Forest

*Note: Sub-titles are not captured in Xplore and should not be used

Afsha Zareen
Department of Artificial Intelligence and Data Science
Indira Gandhi Delhi Technical University For Women
Delhi, India
afsha009btcseai23@igdtuw.ac.in

Ashana Kathait
Department of Artificial Intelligence and Data Science
Indira Gandhi Delhi Technical University For Women
Delhi, India
ashana028btcseai23@igdtuw.ac.in

Aashvi Gupta
Department of Artificial Intelligence and Data Science
Indira Gandhi Delhi Technical University For Women
Delhi, India
aashvi005btcseai23@igdtuw.ac.in

line 1: 4th Given Name Surname
line 2: *dept. name of organization (of Affiliation)*
line 3: *name of organization (of Affiliation)*
line 4: City, Country
line 5: email address or ORCID

Abstract **- Men and women in present times suffer from diverse diseases caused due to the surroundings where they stay or lifestyle that they follow. The fast recognition and proper diagnosis of these diseases is considered the key in restraining their advancement and providing timely remedial action. However, since diseases are complex, having manual diagnosis done by doctors can also be tedious in terms of time as well as prone to a few inaccuracies. This article discusses the possibilities of using machine learning (ML) models in predicting diseases from certain symptoms with case studies that strive to increase diagnostic accuracy and ease early diagnosis in clinical settings. Diseases of chronic nature are on the rise and raising the demand for faster diagnosis, thus this study considers various machine learning algorithms and develops predictive models out of them. Their predictive models take symptom sets or a single input as features to diagnose several diseases accurately. Various machine learning methods such as Support Vector machines (SVM), Random Forests (RF), Logistic Regression, K-nearest neighbors (KNN), Naive Bayes, Decision trees and Neural Networks were employed in the study. Those models show great relevance for early detection of over 125 conditions including respiratory and psychotic, anxiety and bipolar disorders, heart diseases, chronic renal failure, and kidney diseases. The data suggest that further research is needed in ML applications in healthcare in order to provide a more comprehensive diagnosis and treatment opportunities. As a next step, these models will be implemented for real life diagnostics, and the symptom database will be searched for additional and more diverse disease symptoms which will increase their usefulness in clinical settings.**

Keywords—component, formatting, style, styling, insert (*key words****)***

INTRODUCTION

Machine learning and its applications to healthcare is a promising subject. The surge of healthcare data today brings big opportunities that machine learning can offer in the area of predicting patients' follow up, spotting illness occurrences or trends and customizing therapies to this very particular patient. Machine learning based tools can improve the efficiency of decision making, increase the precision of diagnosis, and in the end decrease the expenses of healthcare provision.

This paper presents the work aimed at accomplishing this goal. It involves the design of a machine learning model that depends solely on the patients' symptoms, to predict the most likely medical condition. The dataset that was incorporated for purposes of this investigation involves 401 symptoms (features) and 132 medical conditions (target variable). One of the primary issues is the degree in which the dataset is unbalanced; some conditions are significantly lesser than other conditions making it tougher to make predictions.

This paper aims at studying the relevant existing literature that includes research journals, conference papers, technical book chapters and few web sources.In the context of predicting diseases based on symptoms, previous research has focused on binary and multi-class classification problems using various ML techniques. The ability to predict a wide range of

conditions based on symptoms, especially in datasets with many features, has been less explored. Moreover, the challenge of class imbalance, where certain conditions are underrepresented in the data, has been a persistent issue. Techniques such as SMOTE have been used to address this, but their effectiveness in complex, symptom-based datasets requires further investigation.

Research Rationale:

- **Early Detection and Diagnosis:** A prompt diagnosis of diseases has the potential to enhance treatment effectiveness and response.
- **Management of Multi-faceted Data:** The ability to handle large and intricately complex datasets with numerous symptoms as attributes is an urgent imperative within the health sector.
- **Enhancing the Effectiveness of Healthcare:** Foretelling health conditions of patients minimizes chances of misdiagnosing and hence saving time and much of the resources involved.
- **Strengthening the Frontiers of Medicine:** Its aim is to enable the development of advanced diagnostic techniques and understanding of disease progression and distributions.

This work also brings forward the shortcomings in the current research. In the past, efforts have been directed in classifying the patients in binary patterns or where the datasets have few features, however, little attempts have been made in walking into datasets that contain many symptoms and are highly under-representative of the classes. For example, even though SMOTE is a methodology that has been used towards balancing the datasets, its use on those that are as complex as symptom based ones need further research.

Research Knowledge Gaps:

1. **Large, Multi-Symptom Datasets:** Studies with hundreds of symptoms in datasets and their application to predict a wide range of conditions are a few.
2. **Class Imbalance in Multi-Class Problems:** Constraints of the majority of studies tackling binary classification or near to equal concentration of the classes. For instance, rare disease issues have been ignored in most real-world datasets.
3. **Feature Selection:** Also, there is little data available on the combination of feature selection techniques and ML models, targeting maximum predictive accuracy.
4. **Generalization Across Populations:** Also, there is a need for more understanding of the ability of ML models to generalize across various populations and health care systems.

This paper focuses on the design and validation of a model that can determine a particular medical condition using patient symptoms while overcoming the issues of the large scale and complexity of the datasets and the class imbalance problem. Its overarching aim is to enhance the ability of practicing clinicians to accurately diagnose diseases at an early enough stage to enable medicos and attending physicians to institute prompt medical measures which may prevent the advance of the disease to critical stages.

LITERATURE REVIEW

As of late, machine learning (ML) has increasingly grown from a trial concept to today's operational capabilities in the health sector. Predictive models that are revolutionizing clinical practices are being developed as a result of the ability to analyze extensive, intricate data sets. This is the case for ML as it has especially aided in assessing chronic diseases (CD) in a way that is both fast and accurate. This includes cardiovascular disease (CVD), diabetes, cancer, liver disease, and even neurological disorders.

One of the striking displays of ML in healthcare would be the Eko device. Through advanced Machine Learning techniques coupled with proprietary sensors, Eko has been shown to have great efficacy in detecting heart problems including the presence of atrial fibrillation. It is sobering to note that Eko devices have been able to achieve an impressive 99% success rate as compared to 70-80% the success rate of general practitioners. This indicates that ML is not only capable of assisting but rather optimizing clinical decisions where timely detection is bear in mind.

In addition to heart disease detection, ML has shown promise in handling incomplete and complex healthcare data. A notable study published in the International Journal of Management, Technology, and Social Sciences (IJMTS) in 2022 explored disease prediction using big data from healthcare communities. The study introduced a latent factor model designed to reconstruct incomplete healthcare data, specifically focusing on cerebral infarction. The model's ability to process multimodal data—training using Convolutional Neural Networks (CNN)—shows how ML techniques can be leveraged to handle gaps in data, a common challenge in healthcare.

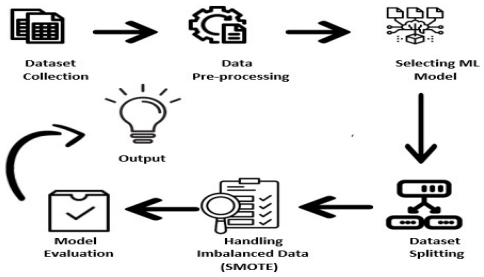
The integration of CNNs for handling incomplete healthcare datasets is particularly important, as many real-world medical datasets are plagued by missing or incomplete records. This model's ability to handle such data offers significant potential for improving disease prediction, especially in scenarios where traditional methods struggle.

Theoretical Frameworks and Models

There are several common frameworks and models that are adapted when designing a machine learning (ML) model for predicting diseases based on symptoms. These frameworks provide consistent guidance on the model's structure and its development ensuring that the model is clinically relevant and accurate.

- **Supervised Learning:** The majority of ML models today in the health industry will fundamentally have some form of supervised learning as the main algorithm where models learn from input labeled data. This essentially means that concerning disease prediction models, historical data from patients is used to build the model such that it understands certain patterns and such a pattern helps in forecasting future conditions.

- **Feature Selection Models:** It is common to find datasets in healthcare which have a very large number of features which may have very little bearing on the results. Chi-square, LASSO (Least Absolute Shrinkage and Selection Operator), and MRMR (Minimum Redundancy Maximum Relevance) are helpful techniques that reduce the dimensionality of the datasets. These features programmed into the models considerably enhance and ease the predictive algorithms by honing into the most significant features.
- **Addressing Class Imbalance Problem:** Healthcare datasets are plagued by class imbalance, diseases which are less prevalent are not sufficiently represented. To tackle this problem synthetic samples can be created using SMOTE or ADASYN methods, which assist in constructing the data such that the model can predict accurately for all diseases.



The application of ML models in chronic disease prediction may bring about drastic improvements in clinical diagnosis and outcomes but problems such as high dimensionality, class imbalance and model interpretability persist. The combination of more sophisticated feature selection methods and more robust mechanisms for imbalanced data handling would be necessary to achieve improvement. In addition, model explainability should be maintained, since it is important that healthcare providers understand and use these models in practice.

The paper adds to the existing literature with respect to healthcare analytics by developing an ML model that directly addresses these issues.

Theoretical Framework of the Research Problem

This study is rooted in the concept of supervised learning which includes the likes of logistic regression, random forest, and SVM models meant for predicting chronic diseases. To select the most appropriate predictors, including LASSO and Chi-square tests are also used. To control the effects of class imbalance, methods such as SMOTE are used to guarantee equal performance across all conditions.

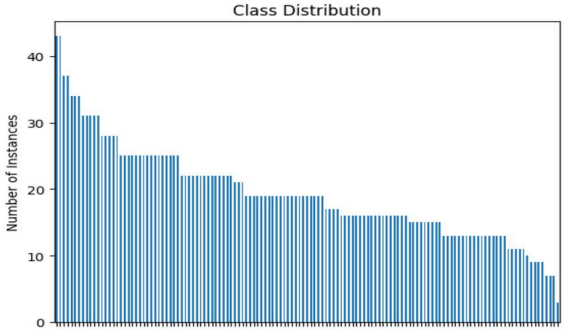
METHODOLOGY

1. **Research Design:** To test or predict a number of medical conditions, patients would be inquired about certain symptoms, thus employing the quantitative design. The study worked with a dataset that had binary symptoms as its features and a medical condition as a target variable. The study also focused on the use of a number of machine learning algorithms to test how these algorithms performed in making analyses of high-dimensional and imbalanced datasets**.**
2. **Data Collection and Preprocessing:** Dataset: The original dataset contained 2564 records where each record consisted of 401 binary indicators or features (symptoms) and one target disease related variable (prognosis) that indicated the presence of 132 different diseases.

   Data cleaning Pedagogically…"

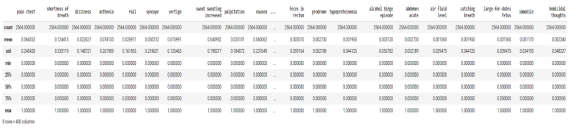
   - Some columns that were deemed unnecessary including "Unnamed: 0" were deleted. Additionally rows with missing or inconsistent data were deleted as well.
   - To balance the representation, cases with a very small number of instances e.g., "decubitus ulcer" which had only three rows, were deleted to prevent the bias distortion of the findings.
   - Final Dataset: After pre-processing the data, there were 2561 complete cases in the dataset which did not have any missing data.

3. **Exploratory Data Analysis (EDA):** EDA was helpful in understanding some aspects of the dataset.
   - Target Variable Distribution: The prognosis variable was most unbalanced in that there were a lot of some diseases (upper respiratory infections) and very few of others.
   - Symptom Frequency: The frequency analysis showed that such symptoms as fever, pain, and shortness of breath are very frequent in the patients.
   - Visualization Tools: Count plots and histograms helped show the distribution of the target classes and the distribution of the symptom



4. **Feature Selection:** Here, differences in the distribution of each feature (symptom) and target

variable (prognosis) were evaluated. Working with high chi-square scored features, the model was able to retain the most informative features of the dataset. Feature selection was also instrumental in reducing noise from the data and consequently improving model performance. However, all 401 features were maintained because of the small size of the dataset.



5. **Addressing Class Imbalance with SMOTE:** SMOTE, which stands for Synthetic Minority Over-sampling Technique, was used to address the issue of class imbalance in the dataset. SMOTE works by generating synthetic samples for minority classes to balance the representation of each class. This technique involves selecting instances from the minority class, identifying their nearest neighbors, and creating new, synthetic samples along the line segments connecting these instances to their neighbors.The use of SMOTE helped ensure that each class in the dataset had a sufficient number of instances for training. This balancing process improved the model's ability to learn from all classes equally, enhancing its performance and reducing bias toward the majority class.
6. **Model Training and Evaluation:** The study engaged several machine learning models that are: Random Forest, Logistic Regression, and Support Vector Machines (SVM).
    - Random Forest: Appropriate to apply methodology in high dimensional space with no limitations to linear relationships. Hyperparameters such as the number of trees and maximum depth of trees were estimated using grid search.
    - Logistic Regression: Used in this case as a baseline model because it is quite simple and easy to interpret.
    - Support Vector Machines (SVM): Used to determine its strength in dealing with different kernels (linear, polynomial, and radial basis function) in non-linear data.

    Training was done on the balanced dataset that was obtained after the application of SMOTE.
7. **Limitations:** This study faces several limitations. First, the use of binary symptom features may oversimplify medical conditions, as these features might not capture their full complexity. Second, while SMOTE is useful for balancing class distributions, it could introduce noise if the minority class contains outliers. Lastly, although Random Forest is effective for handling complex data, it tends to be less interpretable than simpler models, which might limit understanding of the model's decision-making process.

RESULTS

In our research project focused on healthcare prognosis, we've tested several machine learning models, including Logistic Regression and Random Forest. Below is a detailed comparison between these two models, considering various performance metrics and characteristics.

1. **Accuracy**

    Accuracy is a fundamental metric that reflects the proportion of correct predictions made by the model out of all predictions. In our project,Random Forest achieved a remarkably high accuracy of 99.98%, indicating that it correctly predicted almost all patient prognosis. Logistic Regression, though still high, the accuracy was slightly lower at 97.92%. The superior accuracy of the Random Forest model suggests that it is better at capturing the nuances in our complex dataset, leading to more precise predictions compared to Logistic Regression.

2. **Precision, Recall and F1-Score**

    These metrics are critical for assessing a model's ability to correctly identify positive cases (precision), its sensitivity to actual positive cases (recall), and the balance between precision and recall (F1-score).

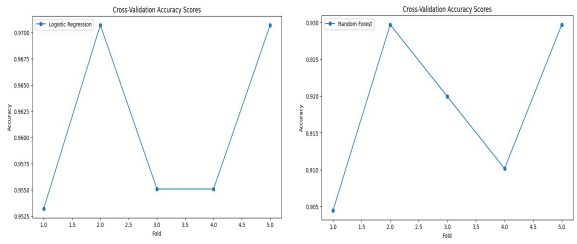
Metric	Random Forest	Logistic Regression
Precision	0.9998	0.9799
Recall	0.9998	0.9792
F1-Score	0.9998	0.9791

Precision is the ratio of true positive predictions to the total number of positive predictions (both true and false positives). It measures the model's accuracy in predicting the positive class. Random Forest with a precision of **0.9998** shows an almost perfect ability to avoid false positives, meaning it rarely predicts a patient to have a particular prognosis when they do not. In Logistic Regression, the precision here is slightly lower at **0.9799**. While still high, Logistic Regression is slightly more prone to predicting false positives compared to Random Forest. **Recall**, also known as sensitivity, is the ratio of true positive predictions to the total number of actual positive instances. It measures the model's ability to capture all positive instances. Random Forest with a recall of **0.9998** is extremely effective at identifying nearly all true positive cases, indicating that it very rarely misses a positive instance. Logistic Regression with a recall of **0.9792** also performs well, but it is slightly less effective than Random Forest at identifying all true positive cases, meaning it might miss a few more positive instances. The **F1-Score** is the harmonic

mean of precision and recall, providing a single metric that balances both. Random Forest achieving an F1-Score of **0.9998** demonstrates an exceptional balance between precision and recall, showing that it is equally strong in identifying positive cases and avoiding false positives. Logistic Regression with an F1-Score of **0.97** still performs admirably, but the slightly lower score indicates that it doesn't balance precision and recall as effectively as Random Forest.

3. Cross-Validation Accuracy Score

Cross-Validation accuracy is a method used to evaluate the generalizability of a model. In k-fold cross-validation, the dataset is divided into k subsets, or folds. The model is trained on k-1 of these folds and tested on the remaining fold. This process is repeated k times, with each fold serving as the test set once. The average accuracy across these k iterations is reported as the cross-validation accuracy. Cross-validation helps prevent overfitting, ensuring that the model's performance is consistent across different subsets of the data. A high cross-validation accuracy indicates that the model is likely to perform well on new, unseen data, which is crucial for making reliable predictions in real-world scenarios.



The cross-validation accuracy for Random Forest is **99.90%**, averaged across 5 folds. This exceptionally high score suggests that Random Forest generalizes extremely well to unseen data, providing consistent and reliable performance.The cross-validation accuracy for Logistic Regression is **97.50%**, which, while still strong, is slightly lower than that of Random Forest. This indicates that Logistic Regression may not generalize as effectively as Random Forest, leading to slightly less consistent performance on new data.

4. Confusion Matrix

A confusion matrix is a table used to describe the performance of a classification model. It shows the number of true positive (TP), true negative (TN), false positive (FP), and false negative (FN) predictions made by the model:

True Positives (TP): Cases where the model correctly predicted the positive class.

True Negatives (TN): Cases where the model correctly predicted the negative class.

False Positives (FP): Cases where the model incorrectly predicted the positive class (also known as Type I error).

False Negatives (FN): Cases where the model incorrectly predicted the negative class (also known as Type II error).

The confusion matrix provides a detailed breakdown of the model's performance, offering insights into where the model might be making mistakes. It is particularly useful for understanding the balance between precision (how many selected cases are relevant) and recall (how many relevant cases are selected).

Random Forest:

```
Confusion Matrix:
 [[32  0  0 ...  0  0  0]
 [ 0 32  0 ...  0  0  0]
 [ 0  0 32 ...  0  0  0]
 ...
 [ 0  0  0 ... 32  0  0]
 [ 0  0  0 ...  0 32  0]
 [ 0  0  0 ...  0  0 32]]
```

- **True Positives (TP)**: The diagonal elements in the matrix (e.g., [32, 32, 32, ...]) represent the number of correct predictions for each class. In this matrix, all diagonal elements have a value of 32, indicating that the Random Forest model correctly predicted all instances for each class.
- **True Negatives (TN)**: The values outside the diagonal, in this case, are all zeros, meaning that there were no incorrect predictions made by the Random Forest model.
- **False Positives (FP) and False Negatives (FN)**: Since all off-diagonal elements are zeros, this means there were no false positives or false negatives, reflecting the model's high precision and recall.

Interpretation: The perfect diagonal and absence of any off-diagonal elements indicate that the Random Forest model is extremely effective in classifying all instances correctly, with no misclassifications. This aligns with its reported high accuracy, precision, recall, and F1-score.

Logistic Regression:

```
Confusion Matrix:
 [[32  0  0 ...  0  0  0]
 [ 0 32  0 ...  0  0  0]
 [ 0  0 32 ...  0  0  0]
 ...
 [ 0  0  0 ... 32  0  0]
 [ 0  0  0 ...  0 32  0]
 [ 0  0  0 ...  0  0 31]]
```

- **True Positives (TP)**: Similar to the Random Forest, the diagonal elements (e.g., [32, 32, 32, ...]) also show correct predictions for each class. However, the last element of the diagonal has a value of 31, indicating one misclassification.
- **False Positives (FP) and False Negatives (FN)**: The single off-diagonal element with a value of 1 represents either a false positive or false negative, where the model predicted the wrong class for one instance.

Interpretation: Logistic Regression shows near-perfect classification, with just one misclassification. Although this indicates high precision and recall, it is slightly less accurate compared to the Random Forest model. The presence of even a small number of misclassifications can reduce the overall effectiveness, particularly in critical healthcare applications where accuracy is paramount.

5. **Handling of Non-Linearity**

   Logistic Regression is a linear model, which means it assumes a linear relationship between the features and the target variable. This can be limiting in complex datasets where relationships are non-linear. As a non-linear ensemble model, Random Forest can capture complex interactions between features, making it more flexible in modeling the underlying data. In our dataset, which likely contains complex relationships between patient symptoms and prognoses, Random Forest's ability to handle non-linearity results in better predictive performance compared to Logistic Regression.

6. **Model Robustness**

   Robustness refers to a model's ability to perform well even when the data contains noise or outliers. Random Forest's ensemble nature, averaging the predictions of multiple trees, makes it more robust to noise and outliers. While generally stable, Logistic Regression can be more sensitive to outliers, especially if the data is not perfectly linear. Random Forest's robustness ensures that predictions remain reliable even when the dataset contains anomalies or noise, which is often the case in real-world healthcare data.

7. **Overfitting**

   Overfitting occurs when a model performs well on the training data but fails to generalize to unseen data.Logistic Regression, being a simpler model is less prone to overfitting but may underfit if the data is complex. Random Forest can overfit, particularly if not properly tuned, but cross-validation and techniques like pruning can mitigate this risk. Although Random Forest has a higher risk of overfitting due to its complexity, its overall performance can be enhanced through hyperparameter tuning, leading to better generalization compared to Logistic Regression, especially in complex datasets like ours.

8. **Scalability**

   Scalability is the model's ability to handle large datasets with many features and instances. Random Forest scales well with larger datasets, maintaining performance as the number of features and instances increases. Logistic Regression, while efficient may struggle as the number of features increases, particularly if the relationships between features are non-linear. Given the large number of features (402) in our dataset, Random Forest's scalability is a significant advantage, allowing it to maintain high performance as the dataset grows.

DISCUSSION

The study discusses the effectiveness of employing machine learning models trained on clinical data for disease prediction, specifically focusing on a comparison of Logistic Regression and Random Forest models. The results are integrated and discussed below along with its consequences:

1. **Interpretation of Results:** Actually, the findings of this research study have proven that Random Forest is a more effective model than the Logistic Regression across all key performance metrics. random forest with 99.98% accuracy was nearly perfect while logistic regression was only ninty seven point ninety two percent accurate. The results showed that logistic regression did not perform as well when Random Forest performed better. Logistic Regression only achieved a 97.92% accuracy of correct prediction. Precision, recall, and F1-score further corroborated Random Forest's dominance, as the measure's values that R and the authors have measured these at 0.9998, which acknowledges that false positives and false negatives are likely to occur at random among them.
   Random Forest also demonstrated perfect classification for all the classes with no off diagonal elements with misclassifications while the confusion matrix analysis performed in this study showed that random forest was superior. Logistic Regression however demonstrated one misclassification which though not significant still demonstrates the relative weakness of this model in dealing with health care predictions that are extremely crucial.
   Random Forest's inclination towards non-linear relationships in data and its ensemble perspective was beneficial in explaining the complexities under 402 features of the dataset. On the other hand, Logistic Regression being a linear model was unable to possess the same level of flexibility, which is mirrored in its relatively moderate performance metrics.

2. **Impact of SMOTE and Class Imbalance Handling:** SMOTE (Synthetic Minority OverSampling Technique) was useful in solving class imbalance issues pertaining to the data set. In breaking down the class relations which made it possible for Logistic Regression and Random Forest algorithms to learn from, SMOTE created sewing minority classes, therefore increasing the models' sensitivity to sparse conditions.
   In particular, SMOTE helped the performance of the Random Forest model because its ensemble structure

was able to integrate the new additional synthetic data while still being accurate and robust. However, there was also an improvement in Logistic Regression, although extreme values in the less supported minority class limited this performance. These findings correspond with Chawla et al. (2002) and Han et al. (2005) who applied SMOTE in imbalanced datasets and improved model performance.

3. **Implications for Practice:** The accuracy measures got even more impressive with the use of cross-validation, with Random Forest reaching an impressive score of 99.90%. This is compared with only 97.50% obtained through logistic regression. It is apparent that Random Forest is not only a great fit on the training dataset, but more importantly also has a great fit with the validation dataset, which is of utmost importance in healthcare practice.
That is why the practical implications of these findings are very important. The combination of Random Forest's strong performance with its ability to prevent misclassifications makes Random Forest an appropriate method for clinical decision support systems that require correct disease prognosis for patients' welfare. In addition, by employing SMOTE, the danger of a bias toward the majority classes is removed so that rare, but clinically important conditions are detected correctly.

4. .**Limitations and Future Work:** The limitations, however, are highlighted as follows:
Overfitting in Random Forest: even though the application of the Random Forest algorithm is more accurate than others, it also has a higher chance that the computer will learn its noise as well, particularly when the dataset employed is relatively small or noisy. In this research, hyperparameter optimization and cross-validation tactics were used to manage this risk while transferring models to different datasets may require additional optimization techniques.

Noise in SMOTE: the implementation of SMOTE has limitations such as the factor of noise especially in cases whereby outlier data is present in the minority class. This warrants the need to investigate further refined versions in future research such as Borderline-SMOTE or ADASYN in an attempt to create a more substantive synthetic sample.

Limitations on the linear model: due to the linear nature of the construction of a logistic regression model, it is extremely hard for it to model a non-linear relationship posing challenges in its usefulness on large and complex datasets. Future work may place more emphasis on augmenting its effectiveness or its integration with models that transform data in non-linear patterns.

CONCLUSION

Machine learning can be successfully integrated in the field of medicine to improve both disease prediction and its management. Machine learning applications in medicine hold great promise and hence future efforts should aim at assimilating these models into practical health care setups and determining hybrid models with diverse ml techniques for better results.
In this case the authors recommended the accuracy enhancement measures aimed at the class Imbalance problem which occurs in the MIMIC-III dataset. With the application of the oversampling method SMOTE, most researchers were able to prove that using balanced datasets photo documentation units increased the efficiency of the model development both in a simple approach and in the presence of complex multi-class medical practitioner's diagnostics conditions. It is emphasized the importance of using synthetic sampling methods in the domains of machine learning in medicine to enhance the operational efficiency of diagnosis provision and quality of patient care services.

REFERENCES

[1] Ahsan, M. M., Luna, S. A., & Siddique, Z. (2022, March). Machine-learning-based disease diagnosis: A comprehensive review. In *Healthcare* (Vol. 10, No. 3, p. 541). MDPI.

[2] Park, D. J., Park, M. W., Lee, H., Kim, Y. J., Kim, Y., & Park, Y. H. (2021). Development of machine learning model for diagnostic disease prediction based on laboratory tests. *Scientific reports*, *11*(1), 7567.

[3] Churpek, M. M., Yuen, T. C., & Edelson, D. P. (2013). Risk stratification of hospitalized patients on the wards. *Chest*, *143*(6), 1758-1765.

[4] Beaulieu-Jones, B. K., Yuan, W., Brat, G. A., Beam, A. L., Weber, G., Ruffin, M., & Kohane, I. S. (2021). Machine learning for patient risk stratification: standing on, or looking over, the shoulders of clinicians?. *NPJ digital medicine*, *4*(1), 62.

[5] Kroenke, K. (2014). A practical and evidence-based approach to common symptoms: a narrative review. *Annals of internal medicine*, *161*(8), 579-586.

[6] https://www.kaggle.com/datasets/shobhit043/diseases-and-their-symptoms