

Prediction of chronic diseases based on ML packages using spark MLlib

Aicha Oussous, Abderrahmane Ez-Zahout, Soumia Ziti

Department of Computer Science, Faculty of Sciences, Mohammed V University in Rabat, Rabat, Morocco

Article Info

Article history:

Received Jun 10, 2024

Revised Sep 20, 2024

Accepted Sep 30, 2024

Keywords:

Apache spark

Breast cancer disease

Chronic diseases

Diabetes disease

Heart disease

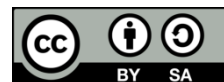
Random forest algorithm

SparkMLlib

ABSTRACT

Heart disease, diabetes, and breast cancer pose significant global health challenges, and effectively addressing these chronic diseases necessitates a coordinated international effort. The integration of machine learning and predictive analytics offers promising solutions for tackling these issues. Our study presents a unified model that utilizes the random forest (RF) algorithm and SparkMLlib to predict these three diseases, testing the model on three distinct datasets and evaluating its performance using scientific metrics, including the receiver operating characteristic (ROC) curve, accuracy, precision, recall, and F1-score. Furthermore, we aim to investigate whether variations in medical data and contextual factors impact the results. The findings indicate that while the model shows strong overall performance, its effectiveness may differ for each disease due to factors such as data characteristics, disease-specific features, model behavior, and various biological and medical considerations; understanding these factors is essential for improving model performance and ensuring its appropriate use in clinical environments.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Aicha Oussous

Department of Computer Science, Faculty of Sciences, Mohammed V University in Rabat

Rabat, Morocco

Email: aicha_oussous@um5.ac.ma

1. INTRODUCTION

Heart disease, diabetes, and breast cancer are significant global health challenges. Heart disease is the leading cause of death, with 17.9 million fatalities in 2016 [1]. Diabetes affected 463 million adults in 2019, projected to rise to 700 million by 2045 [2]. Breast cancer is the most common cancer among women, with 2.3 million new cases diagnosed in 2020 [3]. Heart disease arises from various factors, including genetics and lifestyle, with prevention strategies like lifestyle changes and medications [4]. Diabetes, primarily type 2, results from insufficient insulin production or usage, leading to complications such as cardiovascular issues [5]. Breast cancer treatment varies by type and stage; with early detection improving outcomes. Challenges in addressing these diseases include limited healthcare access in low-income regions and high treatment costs. Unhealthy lifestyles further contribute to the prevalence of heart disease and diabetes [6], [7].

Recent journal articles have explored machine learning techniques for predicting heart disease, diabetes, and breast cancer. Kadhim and Radhi [8] evaluate various algorithms, including support vector machines (SVM), K-nearest neighbors (KNN), decision trees (DTs), and random forest (RF), for classifying heart disease using a dataset from IEEE data port, ultimately finding that RF achieves the highest accuracy at 95.4%, while [9] proposes a model utilizing k-modes clustering and multiple machine learning techniques, achieving accuracy scores between 86.37% and 87.28%, with the multilayer perceptron performing best; [10], the authors aim to construct an efficient model for predicting coronary heart disease (CHD) using seven

algorithms, including RF, which achieved an average accuracy of 92.85%, demonstrating its potential to reduce diagnostic costs and time constraints in medical applications. Kangra and Singh [11] compare various machine learning algorithms, including SVM, Naïve Bayes (NB), KNN, RF, logistic regression (LR), and DT, for predicting diabetes mellitus using the Pima Indian Diabetes (PID) and Germany Diabetes datasets, finding that SVM achieved the highest accuracy of 74% on the PID dataset, while KNN and RF excelled with 98.7% accuracy on the Germany dataset; [12], they focus on identifying crucial features for diagnosing diabetes, determining that the correlation attribute evaluator method is optimal for feature selection and the multiclass classifier is the most effective classifier; finally, [13] presents a model that employs data balancing techniques using SMOTE and various algorithms, with RF yielding the best results at 97% accuracy on the Diabetes dataset 2019 and 80% on the PID dataset, significantly reducing false-negative detections. Nemade and Fegade [14] compare various machine learning classification techniques, including NB, LR, SVM, KNN, DT, and ensemble methods like RF, Adaboost, and XGBoost on a breast cancer dataset, finding that DT and XGBoost achieved the highest accuracy of 97% and an area under the curve (AUC) of 0.999 for XGBoost; in [15], they propose a comparative analysis of eight classification models, identifying SVM as the top performer with an accuracy of 97.7% after applying five feature selection methods; [16] focuses on classifying breast cancer using XGBoost, RF, LR, and KNN, with XGBoost yielding the best results in recall, precision, accuracy, and F1-score; finally, [17] develops a breast cancer risk prediction model using various features, demonstrating that the RF method achieves the highest accuracy of 99.26%, precision of 99%, and AUC of 99%, highlighting the significant impact of multifactorial features on breast cancer risk.

Based on a review of existing solutions, we observed that integrating machine learning and predictive analytics offers promising solutions for addressing chronic diseases [18], [19] and the RF algorithm yields superior results. Our contribution lies in proposing a unified model that utilizes the RF algorithm and SparkMLlib [20], [21] to predict three diseases: heart disease, diabetes, and breast cancer. We will test the model on three distinct datasets and evaluate its performance using scientific metrics such as ROC curve, accuracy, precision, recall, and F1-score. Additionally, we aim to determine whether medical data and context variations impact the results. This paper is structured as follows: section 2 outlines the methods and materials used. In section 3 analyzes the experimental results comprehensively. Finally, section 4 presents the paper's conclusion.

2. METHOD

Our study utilizes three datasets to build prediction models for heart disease, diabetes, and breast cancer, providing detailed information about their structure and the preprocessing steps employed: Cleveland heart disease dataset 2016 [22] comprises 13 independent variables and a class label with five values, which are amalgamated into a binary classification problem; the Vanderbilt-derived diabetes dataset [23] from a study on rural African Americans in Virginia consists of 15 independent variables and a binary class label; and the Breast Cancer Coimbra dataset [24] includes 10 quantitative predictors and a binary dependent variable, with the potential to serve as biomarkers for breast cancer. Data cleaning [25] involved removing irrelevant columns, eliminating duplicate rows, replacing anomalous values, and excluding ID and age attributes, while data preprocessing [26], [27] utilized vector assembler to transform features into a vector format, ultimately enhancing the quality and insights derived from the datasets for the diagnostic models. Table 1 illustrates the features information and description of the heart, diabetes, and breast cancer dataset.

We incorporate essential concepts relevant to this study, including Apache Spark, Spark MLlib, and various machine learning algorithms. Apache Spark [28], [29] is an open-source big data framework designed for the rapid processing of large datasets, capable of handling both structured data, like CSV files, and unstructured data, such as JSON format [30], [31]. A key feature of Apache Spark is its MLlib API [32], which is Spark's machine learning library that provides a variety of algorithms for classification and regression, as well as feature transformations including standardization, normalization, and hashing, along with model evaluation and hyperparameter tuning [33]. In our research, we utilized the MLlib API to develop the offline model component and to implement and assess the RF classification algorithm, using the binary classification evaluator class from the API for evaluating the binary classification models. Additionally, the ML package offers a more recent library of machine learning routines that provides an API for constructing pipelines with data transformers, estimators, and model selectors [34]. This facilitates the seamless integration of multiple data processing and machine learning steps, ensuring consistent application of transformations to both training and new data, thereby simplifying the overall model development, testing, and deployment process. RF is a widely used machine learning classifier for creating predictive models across various research domains [35]-[37], consisting of multiple trees built from randomly selected training datasets and subsets of predictor variables, typically yielding higher accuracy than a single DT mode [38], and will be extensively utilized in this study.

Our model is designed to predict the presence or absence of diseases such as diabetes, breast cancer, and heart disease based on health metrics, utilizing PySpark, a distributed data processing library for Apache Spark, to execute these tasks; the workflow of our study is illustrated in Figure 1.

- Data loading: the model reads health-related features from a CSV file into a Spark DataFrame.
- Data preparation: it explores the data structure, prints the schema, and displays the first few rows.
- Feature vectorization: relevant health features are assembled into a consolidated feature vector for each data point.
- Data splitting: the dataset is partitioned into 80% for training and 20% for testing.
- Model selection: a RF classifier is chosen due to its robustness in managing imbalanced datasets, where one class dominates the other.
- Training the model: the RF model learns patterns and relationships between input features and the target disease classification using the training data.
- Model evaluation: the trained model is assessed using the testing set, with the area under the ROC curve as the evaluation metric [39], [40], recall, accuracy, precision, and F1-score.
- User input for prediction: a function allows users to input their health features for disease prediction.
- Prediction output: the program generates a prediction on the likelihood of having the disease based on the user-inputted health features.

Table 1. Features information and description of heart, diabetes and breast cancer dataset

Attribute heart	Description heart	Attribute diabetes	Description diabetes	Attribute breast cancer	Description breast cancer
AGE	Age	Patient number	Identifies patients by number	Age	
SEX	Sex	Cholesterol	Total cholesterol	BMI	Body Mass Index
CPT	Type of chest pain	Glucose	Fasting blood sugar	Glucose	
RBP	Resting blood pressure	HDL	HDL or good cholesterol	Insulin	
SCH	Serum cholesterol	Chol/HDL	Ratio of total cholesterol to good cholesterol. Desirable result is < 5	HOMA	The homeostasis model assessment (HOMA) employs fasting glucose and insulin levels in the plasma.
FBS	Fasting blood sugar	Age	All adult African Americans	Leptin	A hormone mainly produced by adipose cells that aids in regulating energy balance by reducing hunger
RES	Resting electrocardiographic results	Gender	162 males, 228 females	Adiponectin	A protein hormone involved in the regulation of glucose levels and the breakdown of fatty acids
MHR	Maximum heart rate achieved	Height	In inches	Resistin	ChatGPT A hormone released by adipocytes, known as an adipokine, is associated with obesity and insulin resistance in rodents.
EIA	Exercise-induced angina	Weight	In pounds (lbs)	MCP-1	Monocyte chemoattractant protein-1 strongly attracts monocytes and macrophages to regions of inflammation.
OPK	Old peak = ST depression induced by exercise relative to rest	BMI	$703 \times \text{weight (lbs)} / [\text{height(inches)}]^2$	Classification	Presence or absence of breast cancer disease
PES	Slope of the peak exercise ST segment	Systolic BP	The upper number of blood pressure		
VCA	Number of major vessels (0–3) colored by fluoroscopy	Diastolic BP	The lower number of blood pressure		
THA	Thallium scan	Waist	Measured in inches		
Classification	Presence or absence of heart disease	Hip	Measured in inches		
		Waist/hip	Ratio is possibly a stronger risk factor for heart disease than BMI		
		Classification	Presence or absence of heart disease		

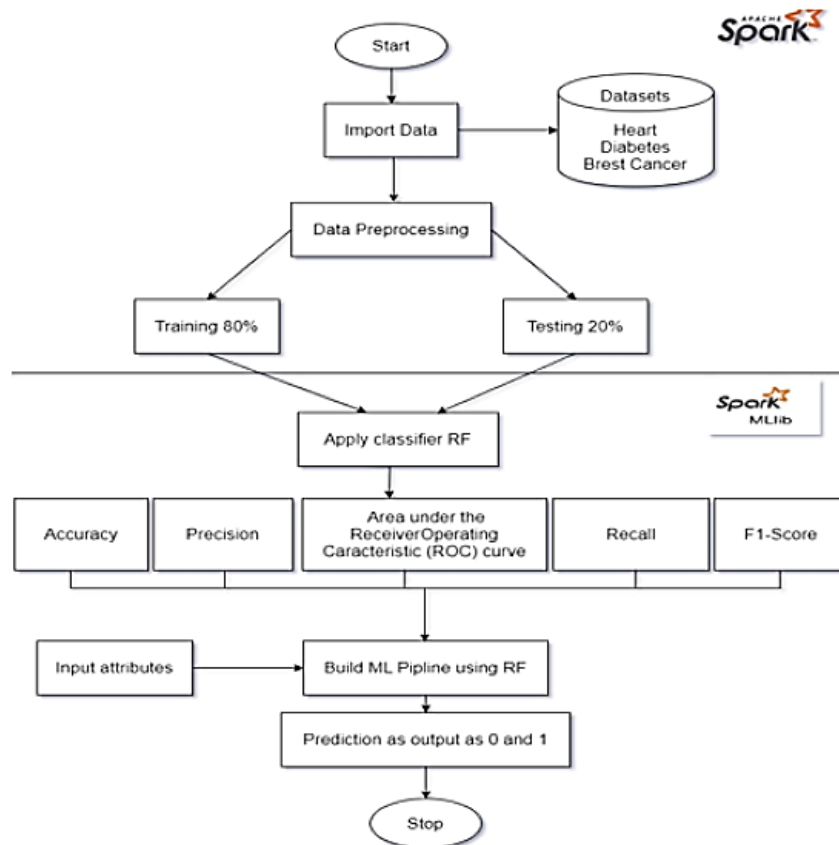


Figure 1. Workflow of the study

3. DISCUSSION AND RESULT

Our visualizations were generated using the Matplotlib and Pandas libraries in Python. The initial visualization features bar charts that depict the likelihood of a disease (such as heart, diabetes, and breast cancer) being absent or present, based on user-inputted features as illustrated in Figures 2-4. The model demonstrates high confidence and accuracy in distinguishing between classes, predicting a 97.99% probability for the absence of diabetes when it is indeed absent, a 97.62% probability for the presence of cancer when it is present, and a 95.69% probability for the presence of heart disease when it is present, with low false positive and false negative rates, highlighting its effectiveness in medical diagnostics.

The second visualization involves constructing a ROC curve to evaluate the classification model. Using the scikit-learn library, the model calculates the false positive rates (FPR) and true positive rates (TPR) to create the ROC curve. The resulting plot illustrates the model's ability to differentiate between classes, with the AUC representing the overall performance for each disease (Diabetes, Heart Disease, and Breast Cancer), as depicted in Figures 5-7. Furthermore, key metrics such as recall, precision, accuracy, and F1-score are computed and presented in Table 2.

The ROC curve illustrates the trade-off between the TPR and the FPR across various thresholds of the model's output. TPR indicates the percentage of true positive cases (individuals with the disease) correctly identified by the model, while FPR indicates the percentage of true negative cases (individuals without the disease) incorrectly classified by the model. A higher TPR and a lower FPR signify a more effective model. The AUC of the ROC curve measures the model's ability to distinguish between positive and negative cases. Ranging from 0 to 1, an AUC of 0 indicates complete inaccuracy, while 1 represents perfection. A random classifier would have an AUC of 0.5, resulting in a diagonal line on the ROC curve. In the provided charts, the machine learning model achieves an AUC of 0.95 for diabetes, 0.86 for breast cancer, and 0.86 for heart disease, indicating a good level of accuracy. The model's ROC curve surpasses the diagonal representing the random classifier for each disease.

This means that the model can achieve a higher TPR than the random classifier at any given FPR. However, the model is not perfect, as it still makes some false positive and false negative errors. The optimal

point on the ROC curve is where the TPR is highest and the FPR is lowest, which corresponds to the best trade-off between sensitivity and specificity of the model.

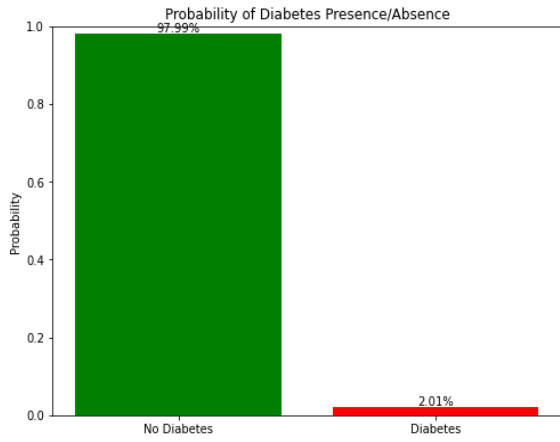


Figure 2. Probability distribution of diabetes

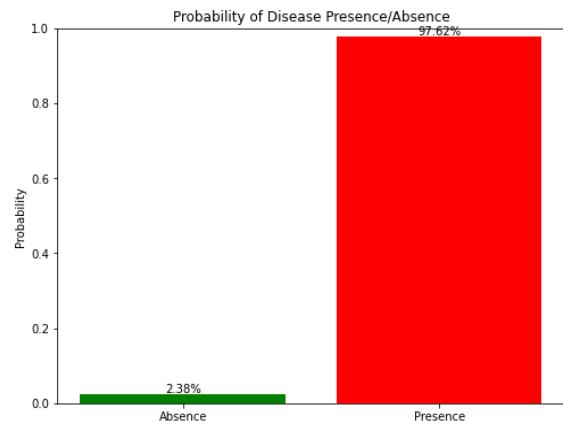


Figure 3. Probability distribution of breast cancer

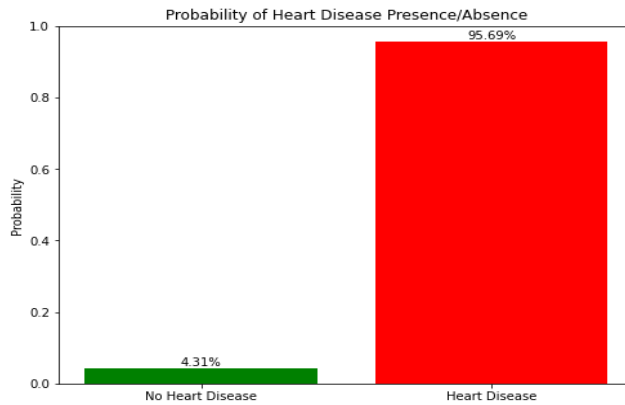


Figure 4. Probability distribution of heart

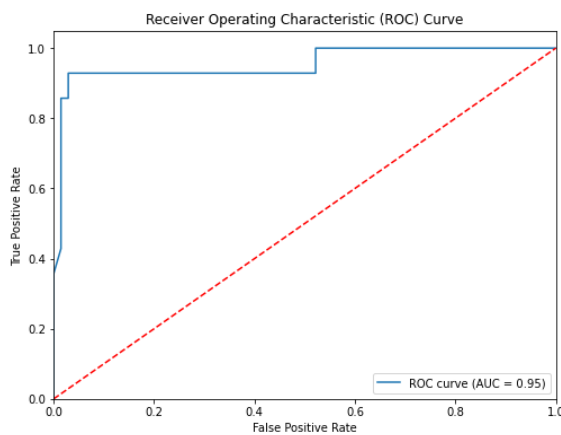


Figure 5. The performance evaluation of diabetes

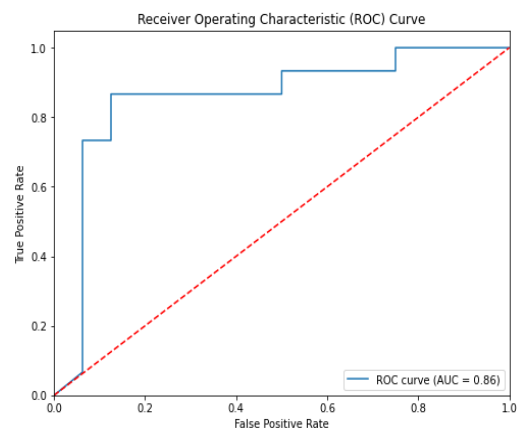


Figure 6. The performance evaluation of breast cancer

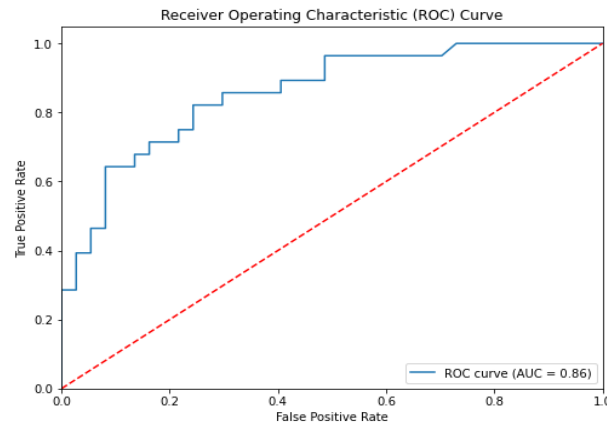


Figure 7. The performance evaluation of heart

Table 2. Result of metrics for diabetes, breast cancer and heart

Diseases	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Diabetes	93.98	90.91	71.43	80.00
Breast Cancer	80.65	76.47	86.67	81.25
Heart	78.46	71.88	82.14	76.67

For diabetes, the model demonstrates excellent overall performance with a high accuracy of 93.98%. The high precision (90.91%) indicates that the model's predictions of diabetes are usually correct. However, the lower recall (71.43%) suggests that the model misses about 29% of actual diabetes cases, aligning with the probability distribution's bias towards predicting no diabetes. The F1-score (80.00%) balances precision and recall, reflecting good overall performance but highlighting the impact of the lower recall.

For Breast Cancer, the model shows good overall performance with an accuracy of 80.65%. The high recall (86.67%) is particularly important in breast cancer detection, as it suggests the model is effective at identifying most cases of cancer, reducing the risk of false negatives. The precision (76.47%) is somewhat lower than the recall, indicating that the model has a tendency to overpredict breast cancer, resulting in more false positives. The F1-score (81.25%) balanced measure of precision and recall indicates good overall performance, aligning well with the accuracy.

For Heart, the model shows good overall performance with an accuracy of 78.46%. The high recall (82.14%) suggests that the model is effective at identifying most cases of heart disease. This is particularly important in medical contexts where missing a positive case (false negative) can have serious consequences. The precision (71.88%) is lower than the recall, indicating that the model has a tendency to overpredict heart disease. This results in more false positives. The AUC of 0.86 indicates that the model has good discriminative ability between the two classes. The F1-score: 0.7667 (76.67%) This balanced measure of precision and recall indicates good overall performance.

Each disease exhibits varying degrees of class imbalance in its dataset. For instance, diabetes has a severe imbalance, with 97.99% of instances representing the "no diabetes" class, while breast cancer and heart disease have less extreme imbalances. This disparity in class representation can significantly impact the model's ability to effectively learn from the minority classes. Moreover, the predictive power of features differs across diseases. Diabetes might have more distinctive features, leading to higher accuracy and AUC scores. In contrast, heart disease could be more complex, with multiple interacting factors, potentially explaining its lower overall performance. Furthermore, each model exhibits varying balances between precision and recall. The breast cancer model favors recall, prioritizing the identification of positive cases, while the diabetes model favors precision, emphasizing the accuracy of positive predictions. The heart disease model strikes a more balanced approach between precision and recall. These differences in performance metrics suggest varying sensitivities to classification thresholds across diseases. Additionally, differences in the amount of available data for each disease could affect model learning and generalization. Diseases with more varied or subtle symptoms, such as heart disease, might be harder to predict accurately compared to diseases with clearer risk factors, like certain types of breast cancer or diabetes. Finally, the same algorithm might perform differently for each disease due to varying data patterns and the level of optimization applied to each model. These factors can collectively influence the overall performance of the models.

In summary, the differences in model performance across various diseases can be attributed to a range of factors, including data characteristics, disease-specific features, model behavior, and various biological and medical considerations. Recognizing these factors is essential for enhancing model performance and ensuring their appropriate use in clinical environments.

4. LIMITATIONS AND FUTURE IMPROVEMENTS

4.1. Limitations

Even though the model demonstrates strong overall performance, it has certain limitations that should be considered: a) Class imbalance: the data exhibits a substantial class imbalance, which could be influencing the model's performance and introducing bias. This skewed probability distribution may lead to inaccurate predictions, particularly for the minority class, b) Overprediction:

The model appears to overpredict the presence of heart disease and breast cancer, which could cause unnecessary worry or lead to additional testing in clinical settings. This tendency to overestimate the likelihood of disease may result in a higher number of false positives. c) Moderate precision: while the model achieves a high recall rate, indicating its ability to identify positive cases, the precision could be improved to reduce the number of false positives. Increasing precision would enhance the model's accuracy in predicting true positives. d) False negatives: the lower recall rate suggests a higher incidence of false negatives, which is particularly concerning in medical diagnostics. False negatives occur when the model fails to identify positive cases, potentially leading to missed diagnoses or delayed treatment. e) Potential overfitting: the extremely high AUC (0.95) combined with the lower recall rate might indicate some degree of overfitting to the majority class. Overfitting happens when a model excels on the training data but struggles to generalize to unseen data, resulting in overly confident predictions and poor performance on real-world scenarios.

4.2. Future improvements

To address these limitations and enhance the model's performance and reliability, we recommend the following improvements: a) Address class imbalance: employ techniques such as synthetic minority over-sampling technique (SMOTE), undersampling, or adjusting class weights to balance the dataset and mitigate the effects of class imbalance. b) Feature engineering: develop more informative features or utilize advanced feature selection techniques to improve the model's ability to capture relevant patterns in the data. c) Threshold adjustment: given the imbalance between precision and recall, adjusting the classification threshold might help strike a better balance between these metrics, depending on the specific requirements of the application. d) Ensemble methods: implement ensemble techniques, like bagging or boosting, which can potentially improve overall performance by combining the strengths of multiple models. e) Deep learning: explore the use of neural networks, as they have the potential to capture complex patterns in the data, particularly for diseases with intricate relationships between features. f) Regularization: implement stronger regularization techniques, such as L1/L2 regularization or dropout, to address overfitting concerns and improve the model's generalization ability. g) Incorporate domain knowledge: collaborate with medical specialists to incorporate domain-specific insights and expert knowledge into the model, which can enhance its performance and interoperability. h) Explainable AI: implement techniques like SHapley additive explanations (SHAP) or local interpretable model-agnostic explanations (LIME) to make the model's decisions more interpretable and transparent. i) Cost-sensitive learning: adopt cost-sensitive learning approaches to penalize false negatives more heavily than false positives, prioritizing the identification of positive cases in medical diagnostics. j) Data collection: collect more diverse and representative data to help balance the dataset naturally and improve the model's ability to generalize to real-world scenarios. By implementing these improvements, the model's performance and reliability can be enhanced, leading to more accurate and trustworthy predictions in clinical settings.

5. CONCLUSION

Our study successfully analyzed data related to heart disease, diabetes, and breast cancer using machine learning techniques such as RF, Pandas, and PySpark. The study found that the model performs best for diabetes prediction (93.98% accuracy, 90.91% precision, and 71.43% recall) and shows balanced performance for breast cancer (86.67% recall) and heart disease (82.14% recall). The model has an AUC of at least 0.80 for all diseases, indicating good performance.

The study suggests that the model could serve as a preliminary screening tool, risk assessment tool, and decision support system in healthcare settings. It also highlights the importance of balancing different performance metrics in medical AI and demonstrates the feasibility of developing multi-disease prediction models with good performance.

However, the study emphasizes the need for careful evaluation and implementation, including extensive clinical validation, ethical considerations, and clear guidelines on integration into clinical practice. Future research should focus on expanding the model to cover more diseases and conditions, investigating its performance on diverse populations, integrating it with electronic health records, and exploring interpretable AI techniques. The study provides a benchmark for future studies in multi-disease prediction models and highlights the potential for earlier disease detection and improved health outcomes.




REFERENCES

- [1] E. J. Benjamin *et al.*, “Heart disease and stroke statistics—2019 update: a report from the american heart association,” *Circulation*, vol. 139, no. 10, Mar. 2019, doi: 10.1161/CIR.0000000000000659.
- [2] “IDF diabetes atlas,” Accessed: Aug. 02, 2024. [Online]. Available: <https://diabetesatlas.org/>
- [3] “Breast cancer,” Accessed: Aug. 02, 2024. [Online]. Available: <https://www.who.int/news-room/fact-sheets/detail/breast-cancer>
- [4] S. S. Virani *et al.*, “Heart disease and stroke statistics—2021 update: a report from the american heart association,” *Circulation*, vol. 143, no. 8, Feb. 2021, doi: 10.1161/CIR.0000000000000950.
- [5] E. D. Parker *et al.*, “Economic costs of diabetes in the U.S. in 2022,” *Diabetes Care*, vol. 47, no. 1, pp. 26–43, Jan. 2024, doi: 10.2337/dci23-0085.
- [6] “Diabetes,” Accessed: Aug. 02, 2024. [Online]. Available: <https://www.who.int/news-room/fact-sheets/detail/diabetes>
- [7] “Cardiovascular diseases (CVDs),” Accessed: Aug. 02, 2024. [Online]. Available: <https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-cvds>
- [8] M. A. Kadhim and A. M. Radhi, “Heart disease classification using optimized machine learning algorithms,” *Iraqi Journal For Computer Science and Mathematics*, pp. 31–42, Feb. 2023, doi: 10.52866/ijcsm.2023.02.02.004.
- [9] C. M. Bhatt, P. Patel, T. Ghetia, and P. L. Mazzeo, “Effective heart disease prediction using machine learning techniques,” *Algorithms*, vol. 16, no. 2, p. 88, Feb. 2023, doi: 10.3390/a16020088.
- [10] A. Hammoud, A. Karaki, R. Tafreshi, S. Abdulla, and M. Wahid, “Coronary heart disease prediction: a comparative study of machine learning algorithms,” *JAIT*, vol. 15, no. 1, pp. 27–32, 2024, doi: 10.12720/jait.15.1.27-32.
- [11] K. Kangra and J. Singh, “Comparative analysis of predictive machine learning algorithms for diabetes mellitus,” *Bulletin EEI*, vol. 12, no. 3, pp. 1728–1737, Jun. 2023, doi: 10.11591/eei.v12i3.4412.
- [12] M. Alzyoud *et al.*, “Diagnosing diabetes mellitus using machine learning techniques,” *International Journal of Data and Network Science*, vol. 8, no. 1, pp. 179–188, 2024, doi: 10.52677/j.ijdns.2023.10.006.
- [13] A. Uddin, M. Islam, A. A. Hossain, A. Akhter, and M. Muntaha, “Machine learning based diabetes detection model for false negative reduction” *Biomedical Materials and Devices*, vol. 2 no. 1, pp. 427-443, 2024.
- [14] V. Nemade and V. Fegade, “Machine learning techniques for breast cancer prediction,” *Procedia Computer Science*, vol. 218, pp. 1314–1320, 2023, doi: 10.1016/j.procs.2023.01.110.
- [15] M. A. Elsadig, A. Altigani, and H. T. Elshoush, “Breast cancer detection using machine learning approaches: a comparative study,” *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 13, no. 1, p. 736, Feb. 2023, doi: 10.11591/ijece.v13i1.pp736-745.
- [16] H. Chen, N. Wang, X. Du, K. Mei, Y. Zhou, and G. Cai, “Classification prediction of breast cancer based on machine learning,” *Computational Intelligence and Neuroscience*, vol. 2023, no. 1, p. 6530719, Jan. 2023, doi: 10.1155/2023/6530719.
- [17] E. Nazari *et al.*, “Breast cancer prediction using different machine learning methods applying multi factors,” *Journal of Cancer Research and Clinical Oncology*, vol. 149, no. 19, pp. 17133–17146, Dec. 2023, doi: 10.1007/s00432-023-05388-5.
- [18] A. Oussous, A. Ez-Zahout, S. Ziti, and A. Oussous, “An overview of the most efficient methods for predicting healthcare disorders,” in *AIP Conference Proceedings*, AIP Publishing, 2023.
- [19] A. Oussous, A. Ez-Zahout, and S. Ziti, “Reviewing chronic ailments: predicting diseases with a multi-symptom approach,” *IJECS*, vol. 35, no. 1, p. 418, Jul. 2024, doi: 10.11591/ijeecs.v35.i1.pp418-427.
- [20] H. Jdi and F. Noureddine, “Precipitation forecasting using machine learning in the region of Beni Mellal-Khenifra,” *Indonesian Journal of Electrical Engineering and Computer Science (IJECS)*, vol. 31, p. 451, Jul. 2023, doi: 10.11591/ijeecs.v31.i1.pp451-458.
- [21] M. Ayoub and F. Algarni, “A healthcare monitoring system for the diagnosis of heart disease in the IoMT cloud environment using MSSO-ANFIS,” *IEEE Access*, vol. PP, pp. 1–1, Jul. 2020, doi: 10.1109/ACCESS.2020.3006424.
- [22] J. Andras, Steinbrunn, William, Pfisterer, Matthias and R. Detrano, “Heart Disease.” 1988.
- [23] “Diabetes Prediction - dataset by informatics-edu,” data.world. Accessed: Jun. 07, 2024. [Online]. Available: <https://data.world/informatics-edu/diabetes-prediction>
- [24] P. Miguel, Pereira, Jos, Crisstomo, Joana, Matafome, Paulo, Seia, Raquel and F. Caramelo, “Breast Cancer Coimbra,” 2018.
- [25] C. Li, *Preprocessing Methods and Pipelines of Data Mining: An Overview*, 2019.
- [26] G. Regulwar, A. Mahalle, R. Pawar, S. Shamkuwar, P. Kakde, and S. Tiwari, “Big data collection, filtering, and extraction of features,” in *Big Data Analytics Techniques for Market Intelligence*, 2023, pp. 136–158, doi: 10.4018/979-8-3693-0413-6.ch005.
- [27] J. Hariharakrishnan, S. Mohanavalli, Srividya, and K. B. S. Kumar, “Survey of pre-processing techniques for mining big data,” in *2017 International Conference on Computer, Communication and Signal Processing (ICCCSP)*, 2017, pp. 1–5, doi: 10.1109/ICCCSP.2017.7944072.
- [28] “Apache Spark™ - Unified Engine for large-scale data analytics,” Accessed: Jun. 08, 2024. [Online]. Available: <https://spark.apache.org/>
- [29] A. Oussous, F.-Z. Benjelloun, A. A. Lahcen, and S. Belfkih, “Big data technologies: a survey,” *Journal of King Saud University - Computer and Information Sciences*, vol. 30, Jun. 2017, doi: 10.1016/j.jksuci.2017.06.001.
- [30] I. Mallahi, J. Riffi, H. Tairi, E.-Z. Abderrahmane, and M. Mahraz, “A distributed big data analytics models for traffic accidents classification and recognition based SparkMLlib cores,” *Journal of Automation, Mobile Robotics and Intelligent Systems*, pp. 62–71, Oct. 2023, doi: 10.14313/JAMRIS/4-2022/34.




- [31] A. Oussous and F. Benjelloun, "A comparative study of different search and indexing tools for big data," *Jordanian Journal of Computers and Information Technology*, vol. 8, p. 1, Mar. 2022, doi: 10.5455/jjcit.71-1637097759.
- [32] X. Meng *et al.*, "MLlib: machine learning in apache spark," *Journal of Machine Learning Research*, vol. 17, May 2015.
- [33] E.-Z. Abderrahmane, "A distributed big data analytics model for people re-identification based dimensionality reduction," *International Journal of High Performance Systems Architecture*, vol. 10, p. 57, Jan. 2021, doi: 10.1504/IJHPSA.2021.10042918.
- [34] P. Hung, T. Hanh, and V. Diep, "Breast cancer prediction using spark MLlib and ML packages," *In Proceedings of the 5th International Conference on Bioinformatics Research and Applications*, Dec. 2018, pp. 52–59, doi: 10.1145/3309129.3309133.
- [35] Md. M. Ali, B. K. Paul, K. Ahmed, F. Bui, J. Quinn, and M. A. Moni, "Heart disease prediction using supervised machine learning algorithms: Performance analysis and comparison," *Computers in Biology and Medicine*, vol. 136, p. 104672, Sep. 2021, doi: 10.1016/j.combiomed.2021.104672.
- [36] P. Ghosh, S. Azam, A. Karim, M. Hassan, K. Roy, and M. Jonkman, "A comparative study of different machine learning tools in detecting diabetes," *Procedia Computer Science*, vol. 192, Jun. 2021, doi: 10.1016/j.procs.2021.08.048.
- [37] R. Shubair, "Comparative study of machine learning algorithms for breast cancer detection and diagnosis," *In 2016 5th International Conference on Electronic Devices, Systems and Applications (ICEDSA)*, Dec. 2016, doi: 10.1109/ICEDSA.2016.7818560.
- [38] L. Breiman, "Random forests," *Machine Learning*, vol. 45, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010950718922.
- [39] A. Linden, "Measuring diagnostic and predictive accuracy in disease management: an introduction to receiver operating characteristics (ROC) analysis. *J Eval Clin Pract* 12: 132-139," *Journal of evaluation in clinical practice*, vol. 12, pp. 132–9, May 2006, doi: 10.1111/j.1365-2753.2005.00598.x.
- [40] M. Khalilia, S. Chakraborty, and M. Popescu, "Predicting disease risk from highly imbalanced data using random forest," *BMC medical informatics and decision making*, vol. 11, p. 51, Jul. 2011, doi: 10.1186/1472-6947-11-51.

BIOGRAPHIES OF AUTHORS






Aicha Oussous    Ph.D. student in Faculty of Sciences, Mohammed V University in Rabat. Computer Science Engineer, a degree from the National School of Applied Sciences, Agadir, Morocco, in 2013. License diploma in mathematics and Computer Sciences from the Faculty of Sciences, Ibn Zohr University, Agadir, Morocco, in 2010. She can be contacted at email: aicha_oussous@um5.ac.ma.



Abderrahmane Ez-Zahout    is currently an associate professor at Department of Computer Sciences/Faculty of Sciences/Mohammed V University. He graduated Ph.D. in Computer Sciences from ENSIAS College of IT. And he is an active member of IPSS team (intelligent processing systems and security). He can be contacted at email: abderrahmane_ezzahout@um5.ac.ma.



Soumia Ziti    is a full professor and researcher at the Faculty of Sciences of Mohammed V University in Rabat since 2007. She obtained her Ph.D. in computer science specializing in graph theory from the University of Orleans in France, along with a diploma in advanced studies in fundamental computer science. She also holds a Baccalaureate in Mathematical Sciences and completed her undergraduate studies in mathematics and physics, specializing in mathematics, at Hassan II University in Morocco. Furthermore, she earned a master's degree in science and technology in computer science from the same institution. Her research interests encompass a wide range of topics including graph theory, information systems, artificial intelligence, data science, software development, database modelling, big data, cryptography, and numerical methods and simulations. Pr. Ziti has contributed extensively to these fields with over than eighty publications in esteemed international journals and conferences. Additionally, she plays a pivotal role in coordinating, participating or assessing in various educational and socio-economic or research projects. She can be contacted at email: soumia.ziti@fsr.um5.ac.ma.