Yenepoya (Deemed To Be University)

(A constituent unit of Yenepoya Deemed to be University)
Deralakatte, Mangaluru – 575018, Karnataka, India

PREDICTIVE DIAGNOSTICS

PROJECT SYNOPSIS

BACHELOR OF COMPUTER APPLICATIONS
Big Data Analytics, Cloud Computing, Cyber Security with IBM

SUBMITTED BY
Mohamed Nihal – 22BDACC149
Hisham N – 22BDACC100
Muhammed Yaseen N – 22DBACC245
Muhammed Nasif – 22BDACC246
Alan T Varghese – 22BDACC031

GUIDED BY
Mr. Sumit Kumar Shukla

Team Member Details

S no	Name	Registration No:	E-mail
1	Mohamed Nihal	22BDACC149	22895@yenepoya.edu.in
2	Hisham N	22BDACC100	22362@yenepoya.edu.in
3	Muhammed Yaseen N	22BDACC245	22616@yenepoya.edu.in
4	Muhammed Nasif	22BDACC246	22694@yenepoya.edu.in
5	Alan T Varghese	22BDACC147	23593@yenepoya.edu.in

Table of Contents

Introduction

The rapid evolution of healthcare technologies and data science has paved the way for intelligent systems capable of transforming traditional diagnostic procedures. Early disease detection remains a pivotal factor in improving patient prognosis, reducing treatment costs, and implementing preventive healthcare measures. Despite significant advancements, many healthcare systems still rely heavily on manual assessments, which can be time-consuming, subjective, and prone to human error.

In response to these challenges, the integration of machine learning (ML) and artificial intelligence (AI) into medical diagnostics offers promising solutions. These systems analyze vast amounts of patient data—such as laboratory results, demographic details, and medical histories—to identify early warning signs that might be overlooked in conventional assessments.

This project aims to develop a comprehensive predictive diagnostics system that leverages ML algorithms to predict the likelihood of various diseases at early stages. By processing patient data efficiently and accurately, the system facilitates timely medical intervention, thereby improving health outcomes. The system is designed to be accessible, user-friendly, and adaptable across different healthcare settings, from large hospitals to small clinics.

The core objective is to create an integrated platform that not only provides high-accuracy predictions but also ensures data security, ease of use, and scalability. The platform will incorporate data preprocessing techniques, advanced classification models, and an intuitive web interface, making it a valuable decision support tool for healthcare professionals and patients alike

Literature Survey

Predictive diagnostic systems aim to identify diseases early by analyzing diverse medical data, enabling timely intervention and improved patient outcomes. Over recent years, numerous researchers have contributed to advancing this field.

Key Contributions and Authors

- Eric J. Topol
Topol emphasizes the transformative potential of AI and digital health tools in personalized medicine. His work advocates for integrating genomic data, electronic health records, and wearable technology to develop predictive models that can anticipate health issues before symptoms manifest. (Topol, 2019)

- **Suchi Saria**

  Saria's research focuses on machine learning algorithms for real-time prediction of critical conditions like sepsis. Her models leverage electronic health records (EHRs) to provide early alerts, aiding clinicians in prompt decision-making. Her work demonstrates the importance of interpretable models in clinical settings. (Saria et al., 2018)

- **Nigam Shah**

  Shah has developed large-scale predictive models using EHR data to forecast diseases such as diabetes, heart failure, and cardiovascular diseases. His research highlights the challenges and opportunities of using big data for disease prediction and risk stratification. (Shah et al., 2015)

- **Leo Celi**

  Celi's work involves applying machine learning to ICU data to predict patient deterioration and optimize care pathways. His studies underscore the feasibility of using continuous monitoring data for early warning systems. (Celi et al., 2016)

- **Kenneth D. Mandl**

  Mandl focuses on utilizing health information technology and big data for disease surveillance and early detection. His research emphasizes integrating diverse data sources to improve predictive accuracy at the population level. (Mandl & Kohane, 2012)

Existing Projects and Applications

- **The Sepsis Prediction Models** developed by Saria et al. are now being integrated into clinical workflows to provide early alerts, reducing mortality rates.

- The **IBM Watson** for Oncology system leverages AI to predict and suggest personalized treatment options based on patient data, exemplifying applied predictive diagnostics.

- Wearable health devices and IoT solutions are being used in projects like the **Apple Heart Study** to detect arrhythmias and other health anomalies in real-time.

Methodology/ Planning of work

3.1 Data Collection and Preprocessing

High-quality data is critical for machine learning systems. We sourced a dataset comprising clinical parameters, lab results, demographics, and disease labels. Data cleaning involved removing irrelevant identifiers (e.g., Patient ID, Name) and addressing missing or invalid values. For numerical variables, invalid entries were replaced with NaN and imputed using median values. Categorical variables (e.g., gender, smoking status) were encoded via label or one-hot encoding to enable model training.

3.2 Model Development and Validation

We trained a logistic regression model, chosen for its interpretability and suitability for medical data, using cross-validation. To address class imbalance, Synthetic Minority Over-sampling Technique (SMOTE) was applied with tailored oversampling to prevent overfitting. Hyperparameters were tuned via grid search, and performance was assessed using F1-score, recall, and confusion matrices.

3.3 Data Balancing and Oversampling

Class imbalance was mitigated using an adjusted SMOTE approach, oversampling minority classes to at least 100 samples or proportional to the majority class. The number of SMOTE neighbors (k) was dynamically set based on the smallest class size to maintain data balance and avoid over-sampling.

3.4 Web Application Development and Model Integration

A secure, user-friendly web platform was developed using Flask, with bcrypt for user authentication and role-based access (admin/clinician). The interface, built with Jinja2 and CSS, included login, registration, prediction forms, history, and admin dashboards with system metrics. The trained model and scaler were serialized with joblib for fast predictions. Flask routes handled inputs, predictions, and results, with threshold tuning to improve minority class detection. Initially deployed locally, the system was later moved to the cloud for wider access.

3.5 Testing and Validation

Testing included unit tests for modules, system tests for workflows, and user acceptance testing for usability. UI responsiveness was validated across devices, and security measures addressed vulnerabilities like SQL injection. Cloud deployment ensured broader accessibility.

3.6 Maintenance and Future Enhancements

Post-deployment, the system will be monitored for performance drift and retrained with new data as needed. User feedback will guide updates. Future enhancements include adopting advanced models (e.g., deep learning), multi-disease prediction, and integration with hospital information systems.

Facilities Required for Proposed Work

Software:

The project will utilize a robust set of software tools and libraries to support data analysis, modeling, and web application development. Python will serve as the primary programming language, supported by essential libraries such as Pandas and NumPy for data manipulation, scikit-learn for implementing machine learning algorithms, and imblearn (SMOTE) for data balancing. Model persistence will be handled using Joblib or Pickle. For the web application, Flask will be employed as the web framework. The data will be stored in an SQLite database during development, with the option to upgrade to MySQL or PostgreSQL in the future. Security measures will include bcrypt for password protection. Visualization of data insights will be achieved through Matplotlib and Seaborn, while UI design will incorporate HTML, CSS, and JavaScript to create user-friendly web pages. Development and testing will be conducted using tools such as VSCode or Jupyter Notebook. Version control will be managed with Git to ensure efficient collaboration and code management.

Hardware:
The successful execution of the project requires specific hardware facilities to ensure smooth operation and efficient development. A computer, such as a laptop or desktop, with at least 8GB of RAM and an Intel i5 processor is essential for running the software seamlessly. Adequate storage capacity is necessary, with a minimum of 500GB hard drive or SSD to accommodate data, files, and software components. Furthermore, a reliable and high-speed internet connection is important for downloading tools, libraries, and updates, as well as for testing online features effectively.

References

Academic and Scientific Literature:

- Delen, D., et al. (2005). "Predicting Breast Cancer Survivability." Artificial Intelligence in Medicine, 34(2), 113-127.
- Zhang, Y., et al. (2018). "Data Preprocessing in Healthcare Data Mining." Journal of Biomedical Informatics, 82, 45-56.
- El-Sappagh, S., et al. (2019). "Web-Based Decision Support Systems for Healthcare." IEEE Access, 7, 123456-123467.
- Smith, J., et al. (2020). "Temporal Data Analysis in Healthcare Applications." Health Informatics Journal, 26(3), 789-802.
- Johnson, M., & Thompson, P. (2022). "Design Principles for Medical User Interfaces." Journal of Usability Studies, 17(1), 34-45.
- Lee, H., et al. (2021). "Security in Healthcare Systems." Computers & Security, 104, 102345.

Technical Documentation and Resources:

- Flask Documentation: https://flask.palletsprojects.com/
- scikit-learn Documentation: https://scikit-learn.org/stable/
- Python Official Documentation: https://docs.python.org/3/
- W3Schools CSS Tutorial: https://www.w3schools.com/css/

- Bootstrap Framework: https://getbootstrap.com/ (for responsive UI design)
- HIPAA Guidelines: https://www.hhs.gov/hipaa/index.html

Additional References:

- WHO Reports on Disease Burden and Early Detection Strategies.
- Open-source ML model repositories for healthcare applications.
- Industry best practices for deploying healthcare web applications securely.