ENHANCED DISEASE PREDICTION USING GENOME-BASED ANALYSIS

A PROJECT REPORT

Submitted by

SNEHA A 312320104143

YAZHINI DEVI G.K.V 312320104183

in partial fulfilment of the award of the degree

of

BACHELOR OF ENGINEERING

in

COMPUTER SCIENCE AND ENGINEERING

St. JOSEPH'S COLLEGE OF ENGINEERING
(An Autonomous Institution)
OMR, Chennai 600 119

ANNA UNIVERSITY :: CHENNAI 600 025

MARCH 2024

ANNA UNIVERSITY, CHENNAI

BONAFIDE CERTIFICATE

Certified that this project report **"Enhanced Disease Prediction using Genome-based Analysis"** is the bonafide work of **SNEHA A (312320104143) AND YAZHINI DEVI G.K.V (312320104183)** who carried out the work under my guidance. Certified further that to the best of my knowledge the work reported herein does not form part of any other thesis or dissertation on the basis of which a degree or award was conferred on an earlier occasion on this or any other candidate.

SIGNATURE

HEAD OF THE DEPARTMENT
Dr.G.Maria kalavathy, M.E., M.B.A., Ph.D.,
Professor & Head of Department
Dept. of Computer Science and Engineering,
St. Joseph's college of Engineering,
OMR,Chennai-600 119

SIGNATURE

SUPERVISOR
Mrs. Jeipratha P N, ME., (Ph.D).,
Assistant Professor,
Dept. of Computer Science and Engineering,
St. Joseph's college of Engineering,
OMR Chennai-600 119

Submitted to Project and Viva Examination held on _____

INTERNAL EXAMINER **EXTERNAL EXAMINER**

ACKNOWLEDGEMENT

At the outset, we would like to express our sincere gratitude to our beloved **Dr. B. Babu Manoharan M.A., M.B.A., Ph.D.,** ***Chairman, St. Joseph's Group of Institutions*** for his constant guidance and support to the student community and the Society.

We would like to express our hearty thanks to our respected***Managing Director*** **Mr. B. Shashi Sekar, M.Sc.** for his kind encouragement and blessings.

We wish to express our sincere thanks to the Executive ***Director***

Mrs. S. Jessie Priya M.Com. for providing ample facilities in the institution. We express sincere gratitude to our beloved **Principal**

Dr. Vaddi Seshagiri Rao M.E., M.B.A., Ph.D., F.I.E. for his inspirational ideas during the course of the project.

We express our sincere gratitude to our beloved Dean ***(Research)***
Dr. A. Chandrasekar M.E., Ph.D., Dean (Student Affairs)
Dr. V. Vallinayagam M.Sc., M.Phil., Ph.D., and Dean (Academics)
Dr. G. Sreekumar M.Sc., M.Tech., Ph.D., for their inspirational ideas duringthe course of the project.

We wish to express our sincere thanks to **Dr.G.Maria kalavathy, M.E., M.B.A., Ph.D.,*Professor and Head of the Department***, Department of Computer Science and Engineering, St. Joseph's College of Engineering for her guidance and assistance in solvingthe various intricacies involved in the project.

We would like to acknowledge our profound gratitude to our supervisor **Mrs. Jeipratha P N, M.E., (Ph.D).,** for her/his expert guidance and connoisseur suggestion to carry out the study successfully.

Finally, we thank the **Faculty Members** and **our Family**, who helped and encouraged us constantly to complete the project successfully.

ABSTRACT

In recent years, the integration of genomic data into medical research has revolutionized disease prediction and diagnosis. This project aims to leverage the vast amount of genetic information available to predict diseases more accurately through a comprehensive genome-based analysis approach. The core idea revolves around utilizing protein IDs as inputs and employing advanced computational techniques to deduce the associated disease with higher precision.

While the proposed research presents an innovative approach to gene prioritization and disease prediction, several limitations should be considered. Firstly, the utilization of multiple techniques such as Recurrent Neural Network (RNN), Deep Belief Network (DBN), fuzzy logic, and an ensemble classifier could significantly increase computational complexity and cost, potentially limiting practical applicability. Secondly, the effectiveness of the model heavily relies on the quality and availability of data, raising concerns about data dependency and generalizability.

The significance of this project lies in its potential to enhance disease prediction accuracy, thereby enabling early detection and proactive intervention strategies. The methodology involves several key steps. First, we gather a comprehensive dataset comprising protein IDs and their corresponding disease annotations. Next, algorithms like KNN, Decision Tree and Naïve Bayes are employed to analyze the genomic sequences associated with each protein ID. Machine learning techniques are then applied to learn complex patterns and relationships within the genomic data, enabling the development of robust disease prediction models.

KEYWORDS: Genomic Data, Protein ID, Machine Learning, Deep Learning, Chronic Disease

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF ABBREVIATIONS

ML	**M**achine **L**earning
KNN	**K**-**N**earest **N**eighbors
SNP	**S**ingle **N**ucleotide **P**olymorphisms
ANN	**A**rtificial **N**eural **N**etwork
AGDPM	**A**dvanced **G**enome **D**isorder **P**rediction **M**odel
TWAS	**T**ranscriptome **W**ide **A**ssociation **S**tudies
FLR	**F**our **L**ist **R**epresentation
IIOT	**I**ndustrial **I**nternet **o**f **T**hings
AD	**A**lzheimer's **D**isease
IDE	**I**ntegrated **D**evelopment **E**nvironment
ORM	**O**bject-**R**elational **M**apping
UML	**U**nified **M**odelling **L**anguage

CHAPTER 1

INTRODUCTION

1. INTRODUCTION

In recent years, the field of medical science has witnessed a remarkable shift towards personalized healthcare, driven by advancements in genomic analysis and bioinformatics. The ability to decode the human genome has opened up unprecedented opportunities to understand the genetic basis of diseases and develop more precise diagnostic and therapeutic strategies. One of the most promising applications of genomic analysis is disease prediction, where the genetic information of an individual is utilized to anticipate their susceptibility to various health conditions. In line with this, the project titled "Enhanced Disease Prediction Using Genome Based Analysis" aims to leverage genomic data to predict diseases with higher accuracy and efficiency.

The traditional approach to disease prediction primarily relies on clinical parameters and demographic information. While valuable, these methods often lack the granularity necessary to capture the intricate geneticpredispositions that underlie many complex diseases. By contrast, genome- based analysis offers a more comprehensive and personalized approach by examining the unique genetic makeup of an individual. With the advent of high-throughput sequencing technologies and advancements in computationalbiology, it is now possible to analyze vast amounts of genomic data efficiently and extract meaningful insights regarding disease susceptibility.

The core concept of the project revolves around utilizing protein identifiers (IDs) as input data to predict the likelihood of various diseases. Proteins play a crucial role in biological processes, and alterations in their expression or function can be indicative of underlying health conditions. By analyzing

protein data in conjunction with genomic information, the project seeks to enhance the accuracy and specificity of disease prediction models. This approach offers several advantages, including the ability to identify subtle genetic variations and biomarkers associated with specific diseases, leading to more targeted and personalized interventions.

Moreover, the project aims to harness the power of machine learning algorithms to analyze complex genomic datasets and generate predictive models. Machine learning techniques, such as deep learning and ensemble methods, have demonstrated remarkable performance in various bioinformatics tasks, including disease prediction. By training these algorithms on large-scale genomic datasets, the project endeavors to develop robust and scalable models capable of accurately predicting disease outcomes based on protein IDs.

1.2 WHAT IS THE NEED OF GENOME-BASED ANALYSIS?

Genome-based analysis allows for a deeper understanding of individual genetic variations, which can influence disease susceptibility, treatment response, and overall health outcomes. By tailoring medical interventions to an individual's unique genetic makeup, precision medicine aims to maximize efficacy while minimizing adverse effects.

Many diseases have a genetic component that can be detected through genome-based analysis even before symptoms manifest. Early detection enables proactive interventions, potentially preventing the progression of diseases or enabling treatment at an early stage when outcomes are more favorable. By analyzing an individual's genome, healthcare providers can assess their predisposition to certain diseases or adverse drug reactions. This personalized risk assessment enables targeted screening programs, lifestyle

modifications, and preventive measures tailored to each individual's genetic profile.

Genome-based analysis plays a crucial role in drug discovery and development. By identifying genetic targets associated with diseases, researchers can develop more effective and targeted therapies, leading to better treatment outcomes and reduced side effects.

Analyzing the genome provides insights into the underlying molecular mechanisms of diseases, shedding light on their pathogenesis and progression. This deeper understanding enhances researchers' ability to develop novel therapies and diagnostic tools.

Genome-based analysis can inform public health initiatives by identifying population-level genetic trends, environmental interactions, and disease risk factors. This information can guide the development of targeted interventions and policies aimed at reducing the burden of disease within specific populations. Genome-based analysis contributes to the advancement of scientific knowledge by uncovering new genetic associations, pathways, and regulatory mechanisms. This knowledge not only enhances our understanding of human biology but also informs future research directions and therapeutic strategies.

1.3 SUPERVISED MACHINE LEARNING

Supervised machine learning is a subset of artificial intelligence (AI) and machine learning techniques where the algorithm learns from labeled data to make predictions or decisions. In supervised learning, the algorithm is provided with a dataset consisting of input-output pairs, where each input is associated with a corresponding output or label. The goal of supervised learning is to learn a mapping or relationship between the input variables and

the output variable based on the provided data. During the training phase, the algorithm iteratively adjusts its internal parameters to minimize the error between its predictions and the actual labels in the training data. Common types of supervised learning algorithms include regression, where the output variable is continuous and predictive modeling aims to predict a numerical value, and classification, where the output variable is categorical and the algorithm predicts which category or class the input belongs to. Supervised learning finds wide applications across various domains, including image and speech recognition, medical diagnosis, financial forecasting, and natural language processing. Its effectiveness relies heavily on the quality and quantity of labeled training data, as well as the choice of appropriate algorithms and model evaluation techniques to ensure accurate and reliable predictions.

1.4 NAÏVE BAYES

Naive Bayes is a popular and fundamental algorithm in the realm of machine learning and probabilistic modeling, known for its simplicity, efficiency, and effectiveness in various applications. At its core, Naive Bayes is a probabilistic classifier based on Bayes' theorem, which provides a principled framework for making predictions using probability theory. The "naive" in Naive Bayes stems from the assumption of independence among the features, meaning thateach feature contributes independently to the probability of a particular class label given the input data. While this assumption may not hold true in all real-world scenarios, Naive Bayes often performs surprisingly well in practice andis widely used in text classification, spam filtering, sentiment analysis, and other tasks.

The key concept behind Naive Bayes is conditional probability, which quantifies the likelihood of an event occurring given that another event has

already occurred. In the context of classification, Naive Bayes calculates the probability of each class label given the input features and selects the label with the highest probability as the predicted class. This is achieved by decomposing the joint probability of the class label and the input features into a product of conditional probabilities using Bayes' theorem. Despite its simplicity, Naive Bayes can effectively handle high-dimensional data and large feature spaces, making it computationally efficient and scalable. Moreover, Naive Bayes is robust to noisy and irrelevant features due to its probabilistic nature. It can handle missing data gracefully by incorporating it into the probability calculations, thereby avoiding the need for imputation or preprocessing steps. Additionally, Naive Bayes requires minimal tuning of hyperparameters, making it easy to implement and deploy in real-world applications, especially when computational resources are limited.

1.5 DECISION TREE

A decision tree is a versatile and intuitive machine learning algorithm used for both regression and classification tasks. Its structure resembles a tree, where each internal node represents a "decision" based on a feature attribute, each branch represents the outcome of that decision, and each leaf node represents the final decision or prediction. Decision trees are popular due to their simplicity, interpretability, and ability to handle both numerical and categorical data.

At the core of a decision tree is the process of recursively partitioning the feature space into smaller subsets based on the values of input features. The algorithm selects the feature that best splits the data into more homogeneous subsets, typically using metrics such as Gini impurity or information gain. The goal is to maximize the homogeneity or purity of the resulting subsets, leading to more accurate predictions. The decision-making process of a decision tree

begins at the root node, where the algorithm selects the feature that best separates the data into distinct classes or categories. It then creates branches corresponding to each possible value of that feature. This process repeats recursively for each subsequent node until reaching a leaf node, where the final decision or prediction is made based on the majority class or average value of the instances within that node.

Decision trees offer several advantages, including their simplicity and interpretability, as the resulting tree structure can be visualized and easily understood by non-experts. They can handle both numerical and categorical data without the need for preprocessing, making them versatile for various types of datasets. Additionally, decision trees implicitly perform feature selection by identifying the most informative features for decision-making.

1.6 K-NEAREST NEIGHBORS

K-Nearest Neighbors (K-NN) is a popular and intuitive supervised machine learning algorithm used for classification and regression tasks. It operates on the principle of similarity: objects that are similar to each other are more likely to belong to the same class or have similar characteristics. K-NN is a non-parametric algorithm, meaning it doesn't make assumptions about the underlying data distribution. Instead, it makes predictions based on the nearest neighbors in the feature space.

In the context of classification, given a new, unlabeled data point, K-NN classifies it based on the class labels of its K nearest neighbors in the training dataset. The value of K, a hyperparameter, determines the number of neighbors considered. To make a prediction, the algorithm calculates the distance between the new data point and each point in the training set using adistance metric such as Euclidean distance, Manhattan distance, or cosine

similarity. The K nearest neighbors are then identified based on these distances, and the majority class among them is assigned to the new data point as its predicted class. One of the key decisions in implementing K-NN is choosing an appropriate value for K. A smaller value of K results in more flexible decision boundaries but may lead to increased sensitivity to noise in the data. Conversely, a larger value of K may lead to smoother decision boundaries but risks oversimplification of the model. The choice of K depends on factors such as the complexity of the problem, the size and distribution of the data, and the desired trade-off between bias and variance.

In regression tasks, K-NN predicts the continuous value of the target variable for a new data point by averaging the values of its K nearest neighbors. This approach assumes that similar data points have similar target values and aggregates them to estimate the value for the new data point. Like in classification, the choice of K in regression tasks influences the smoothness of the predicted function and the model's sensitivity to noise. Despite its simplicity and intuitive nature, K-NN has several limitations. It can be computationally expensive, especially with large datasets, as it requires calculating distances between the new data point and all points in the training set. Additionally, K-NN is sensitive to irrelevant or redundant features and requires careful preprocessing and feature selection to perform effectively. Moreover, the choice of distance metric and the value of K can significantly impact the algorithm's performance and generalization ability.

CHAPTER 2

LITERATURE SURVEY

2.1 EXISTING SYSTEM

The process of prioritizing candidate genes for genome-based diagnostics of hereditary disorders is crucial yet challenging due to the abundance of noisy and specific information surrounding genes, illnesses, and their relationships. While various computer methods for disease gene prioritization have been developed, their effectiveness is often hindered by manual trait creation, limitations in network architecture, or predefined data fusion criteria. To address these limitations, this research proposes a novel gene prioritization and disease prediction model.

Initially, the gathered information undergoes preprocessing through a data cleaning model to ensure accuracy and consistency. In the subsequent gene prioritization phase, the preprocessed data is tokenized and utilized to construct a new knowledge-based ontology structure. This structure incorporates an improved skewness-based semantic similarity function, enhancing the accuracy of gene prioritization. Moreover, an ensemble classifier is formulated, combining Recurrent Neural Network (RNN), optimized fuzzy logic, and Deep Belief Network (DBN) techniques to predict gene disorders effectively.

The ensemble classifier utilizes features extracted from the preprocessing phase for training the RNN, while the constructed knowledge bases train the DBN. Subsequently, the results from both networks are fed into the optimized

fuzzy logic system. The fuzzy logic component plays a pivotal role as the primary indicator, with its fuzzification function fine-tuned using a methodology aimed at improving illness prediction accuracy. Additionally, a newly recommended hybrid system, named Cauchy's Mutated Corona Virus Optimization Algorithm (CMCOA), serves as an upgraded version of the CVOA, a conventional coronavirus optimization technique.

To evaluate the efficiency of the proposed model, a comprehensive comparison is conducted against existing models, considering various performance measures. Notably, the proposed model achieves a remarkable accuracy of 93% with 60% training data, marking significant improvements over existing models such as GCN, SVM, CNN, and various other techniques. Specifically, the precision of the suggested approach with improved features and CMCOA outperforms existing methodologies, highlighting its potential for advancing genome-based diagnostics of hereditary disorders.

2.2 RELATED WORKS

A paper based on gene prioritization and disease prediction model utilizing a combination of semantic similarity functions, ensemble classifiers (including Recurrent Neural Networks and Deep Belief Networks), and an optimized fuzzy logic system was proposed. The model incorporates a hybrid optimization algorithm called Cauchy's Mutated Corona Virus Optimization Algorithm (CMCOA). Comparative evaluation demonstrates significant improvements in accuracy and precision over existing models, achieving a precision improvement of up to 14.42% with the inclusion of CMCOA and improved features.

A study based on the AlexNet Neural Network technique to predict single gene inheritance disorder and multifactorial gene inheritance disorder with respect to various statistical performance parameters to design an advanced genome disorder prediction model (AGDPM) is introduced. AGDPM provided better results than the AlexNet Neural Network in predicting geneticdisorders. This innova- tive model can improve biomedical research and mitigate the high mortality rates associated with genetic disorders.

A study for the prediction of Alzheimer's disease-associated genes to detect the pres ence of Alzheimer's disease is designed. Transcriptome-wide association studies (TWAS) emerge as crucial tools for predicting disease genes by integrating regulatory data and GWAS statistics. To address potential inconsistencies in TWAS analyses using diverse GWAS datasets, anensemble approach was employed. The results upon comparison identified validated genes where AZGP1 surfaced as a potentially associated gene.

An innovative approach to sequence analysis through the innovative Four-List Representation (FLR) of DNA sequences was introduced. The main idea is based on introducing a new representation of the DNA sequence, which breaks the dependency between the DNA bases that exist in the traditional string presentation. It has the ability to generate evidence-based rationales, including similarity maps, scores, and graphs.

A paper leveraging Abstract Meaning Representation is proposed. By employing a data-driven strategy and utilizing convolutional neural networks, the proposed model detects and categorizes the level of profanity in online text content. Compared to commonly developed toxic content detection systems that use lexicon and keyword-based detection, this paper embraces a different approach based on the meaning of the sentence. Meaning representation is a way to grasp the meaning of linguistic input.

A novel approach combining the lasso and the standardized group lasso, mitigating multi- collinearity in linear regression analyses is designed. The proposed fitted sparse-group lasso, implemented using proximal-averaged gra dient descent in the R package “seagull,” prioritizes meaningful weighting of predicted outcomes, particularly crucial in contexts like breeding populations.

A study that emphasizes the importance of Collaborative Intrusion Detection Systems in safeguarding Industrial Internet of Things (IIoT) against cyberattacks is put forward. The paper introduces a device integrity check mechanism based on the concept of a “Digital Genome.” The proposed integrity attestation protocol holds promise for diverse IoT applications.

A study that focuses on Alzheimer’s disease (AD) that utilizes Single Nucleotide Poly- morphisms (SNPs) as crucial biomarkers is designed. This study employs Machine Learning techniques such as Naive Bayes, Random Forest, Logistic Regression, and Support Vector Machine on AD genetic data from the ADNI-1 Whole-genome sequencing datasets. Naive Bayes demonstrated high performance in early Alzheimer’s detection.

A paper to address a multi-assembly problem, where the objective is to reconstruct multiple genomic sequences from mixed reads originating from different sources is proposed. This focuses on constrained path covers, incorporating practical constraints in multi-assembly problems. Efficient algorithms are used for finding all maximalsafe paths.

A paper that addresses the crucial task of genome classification focusing on the utilization of the C5.0 algorithm to highlight its potency in the classification process was introduced. It explores the relationship between key genomicelements-CDS, GC percent, and Size (Mb).

2.3 PROPOSED SYSTEM

The proposed system aims to revolutionize disease prediction by leveraging genome-based analysis techniques. At its core, the system will utilize advanced algorithms and computational methods to analyze genetic data associated with specific protein IDs provided as input. Through this process, the system will be capable of predicting the likelihood of various diseases based on the genetic makeup of an individual. The first step in the proposed system involves the collection and preprocessing of genetic data. This entails gathering protein IDs from individuals, extracting relevant genomic information, and organizing it into a format suitable for analysis.

Once the genetic data is prepared, the system will employ state-of-the-art machine learning and data mining algorithms to perform comprehensive genome-based analysis. This analysis utilizes algorithms like K-Nearest Neighbors, Decision Tree and Naïve Bayes. Django Cloud platform is also used in the backend to process large datasets. By leveraging large-scale genomic databases and incorporating cutting-edge bioinformatics tools, the system will be able to uncover subtle genetic variations and their potential implications for disease susceptibility. To enhance the predictive accuracy of the system, a multi-layered approach will be adopted. This includes integrating various types of genetic data, such as single nucleotide polymorphisms (SNPs), gene expression profiles, and protein-protein interactions, to capture the complexity of biological systems accurately. Additionally, advanced feature selection and dimensionality reduction techniques will be employed to identify the most informative genetic features and mitigate the curse of dimensionality.

Furthermore, the proposed system will leverage ensemble learning methods to combine the predictions of multiple models and improve overall

performance. Ensemble techniques, such as random forests, gradient boosting, and stacking, will be employed to harness the collective intelligence of diverse algorithms and mitigate individual biases and errors. The output of the system will be the prediction of disease likelihood based on the provided protein IDs. This prediction will include not only the name of the disease but also the probability or confidence level associated with the prediction.

CHAPTER 3

SYSTEM REQUIREMENTS

3.1 HARDWARE REQUIREMENTS

System : Pentium IV 2.4 GHz

Hard Disk : 40 GB

Floppy Drive : 1.44 Mb

Monitor : 15 VGA Colour

Mouse : Logitech

Ram : 512 Mb

3.2 SOFTWARE REQUIREMENTS

Python: Python, renowned for its user-friendly syntax and extensive ecosystem, has gained prominence not only for its ease of learning but also for its applicability across diverse domains. In the realm of web development, frameworks like Django and Flask leverage Python's simplicity to streamline the creation of robust and scalable web applications. Furthermore, Python's dominance extends into data science and machine learning, where libraries such as NumPy, Pandas, and TensorFlow empower researchers anddevelopers to efficiently manipulate data and build sophisticated models. Thelanguage's support for multiple programming paradigms allows developers tochoose the approach that best suits their project, contributing to its adaptability. Python's dynamism, while offering flexibility, necessitates thoughtful consideration of variable types during coding to avoid runtime errors. The dynamic typing feature, however, aligns with Python's emphasis on rapid development and prototyping.

Jupyter Notebook: The Jupyter Notebook, born out of the amalgamation of Julia, Python, and R, has become a cornerstone in interactive computing and collaborative research. Beyond its support for multiple programming languages, the open-source web application is renowned for its versatility in creating dynamic documents that seamlessly blend code, visualizations, and textual explanations. The interactive computing environment offered by Jupyter Notebook facilitates an iterative and exploratory approach to coding, making it particularly valuable for data analysis and scientific research. The notebook structure accommodates both code cells and markdown cells, allowing users to intersperse executable code with narrative text, mathematical equations, and visual elements. The output from code cells extends beyond simple text, encompassing diverse media such as images, interactive widgets, and plots generated by libraries like Matplotlib, Seaborn,and Plotly. Jupyter's compatibility with various data visualization tools furtherenhances its utility for creating and showcasing charts and graphs.

Anaconda Navigator: Anaconda Navigator stands as a key component within the Anaconda distribution, catering to the needs of data scientists and researchers engaged in data science and scientific computing. As a graphical user interface (GUI), it offers an intuitive platform for overseeing and launching applications, environments, and packages integral to the Anaconda ecosystem. Anaconda Navigator extends its utility by enabling the creation and management of isolated Python environments, a crucial capability for maintaining dependencies and package versions unique to various projects. The Navigator GUI seamlessly integrates popular integrated development environments (IDEs) like Jupyter Notebooks, JupyterLab, and Spyder, providing users with versatile options for data analysis, machine learning, and scientific computing. Such integration streamlines the development workflow and supports diverse coding preferences. Furthermore, Anaconda Navigator

demonstrates adaptability by incorporating cloud integration features in certain versions.Its emphasis on a user-friendly graphical interface positions it as an accessible and efficient resource in the Anaconda ecosystem, particularly for those who prefer visual interactions over traditional command-line approaches

Django: Django's ORM (Object-Relational Mapping) is a key feature that simplifies database interactions. By defining models as Python classes, developers can seamlessly work with databases without delving into complex SQL queries. This abstraction not only enhances code readability but also ensures database independence, allowing developers to switch between different database backends. One notable aspect of Django is its robust administration interface. This built-in tool empowers developers and administrators to efficiently manage various aspects of the application, such as content, users, and permissions. This eliminates the need for creating a custom admin system, saving time and effort during the development process. In addition to its powerful core features, Django promotes the development of reusable applications. Django is a versatile framework employed for diverse web applications. Its applications span from content management systems and e-commerce platforms to social networking sites and beyond, showcasing its adaptability and scalability in meeting different project requirements.

CHAPTER 4

SYSTEM DESIGN

4.1 SYSTEM ARCHITECTURE

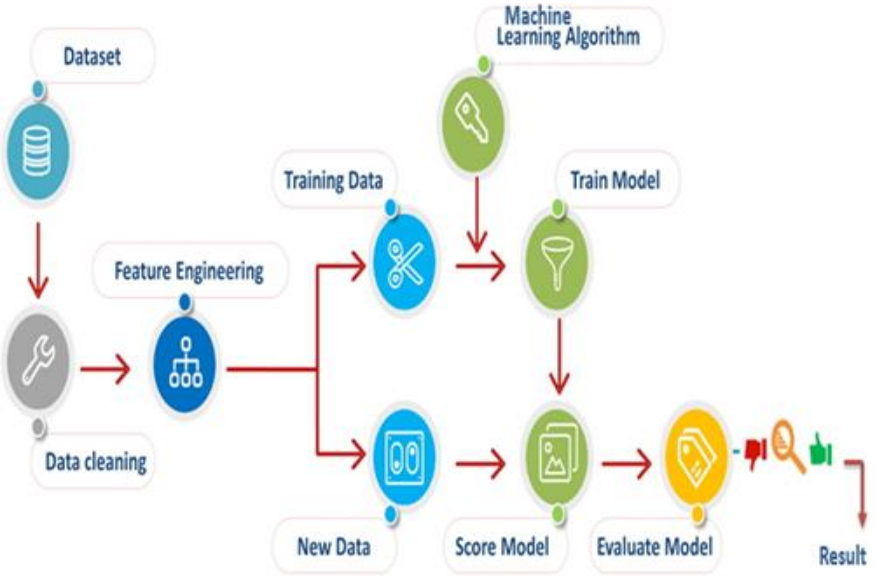

Fig 4.1: Architecture of the System

4.2 UML DIAGRAMS

Unified Modeling Language (UML) is a standardized visual language widely used in software engineering to model and document complex software systems. UML provides a set of diagram types, each serving a specificpurpose, to represent various aspects of a system's architecture, structure, andbehavior. It offers a common visual vocabulary that facilitates communicationamong stakeholders, including developers, architects, and project managers. Unified Modeling Language (UML) diagrams are visual representations used in software engineering to depict the structure, behavior, and interactions 18 within a system. These diagrams serve as a universal language for

communication among software developers, architects, and other stakeholders. Each UML diagram type, such as class diagrams, sequence diagrams, and use case diagrams, addresses specific aspects of a system, allowing for a comprehensive and standardized approach to system design and documentation. UML diagrams help in visualizing the relationships between various components, modeling the flow of activities, and providing a clear understanding of the system's architecture. Their versatility makes them an indispensable tool in the software development process, facilitating collaboration, design, and communication throughout the project lifecycle.

4.2.1 USE CASE DIAGRAM

A Use Case Diagram is a visual representation within the Unified Modeling Language (UML) that illustrates the interactions between actors (external entities) and a system to achieve specific functionalities. It provides a high-level view of the system's functionality from an end-user perspective, outlining the various use cases and their relationships. Actors are entities that interact with the system, while use cases represent specific functionalities or features. In a Use Case Diagram, actors are depicted as stick figures, and use cases are represented by ovals. Lines connecting actors and use cases indicate the interactions. The diagram helps in identifying, clarifying, and organizing system requirements by showcasing how external entities interact with the system to achieve particular goals. Use case diagrams are essential during the early stages of software development to provide a high-level overview of system requirements and functionality. They help stakeholders, including developers, designers, and clients, to understand the system's behavior and the roles of different actors within it.

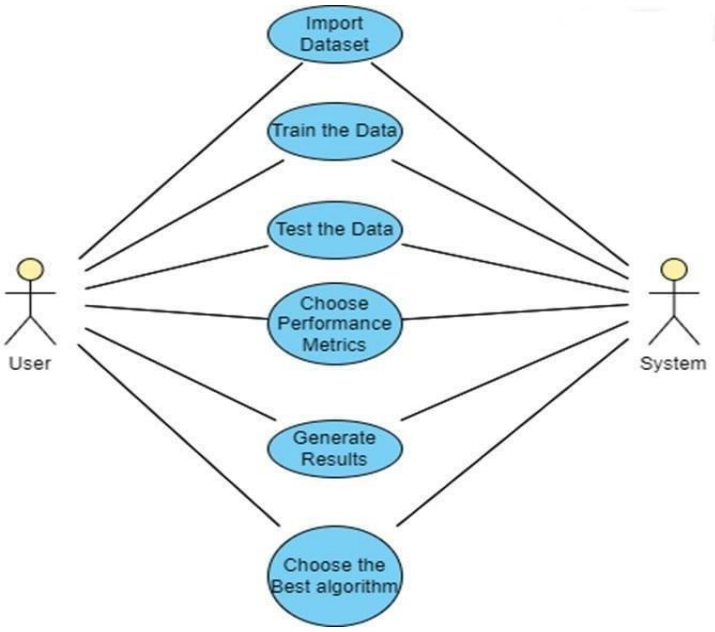

Fig 4.2 Use Case Diagram

4.2.2 SEQUENCE DIAGRAM

A Sequence Diagram is a type of Unified Modeling Language (UML) diagram that illustrates the dynamic interactions and message flows among various objects or components within a system over time. It provides a chronological representation of how different entities collaborate to accomplish a particular functionality. Sequence diagrams are particularly useful for visualizing the order of messages exchanged between objects during the execution of a use case or scenario. In a Sequence Diagram, lifelines represent the objects or participants in the interaction, and vertical lines called activations show when those objects are active during the sequence of events. Arrows and messages indicate the flow of communication between objects. Sequence diagrams are useful for visualizing the dynamic behavior of a system, especially during the design and analysis phases of software development. They provide a clear and concise representation of how objects collaborate to achieve a particular functionality or behavior, making them an essential tool for communication among stakeholders, including developers, designers, and project managers.

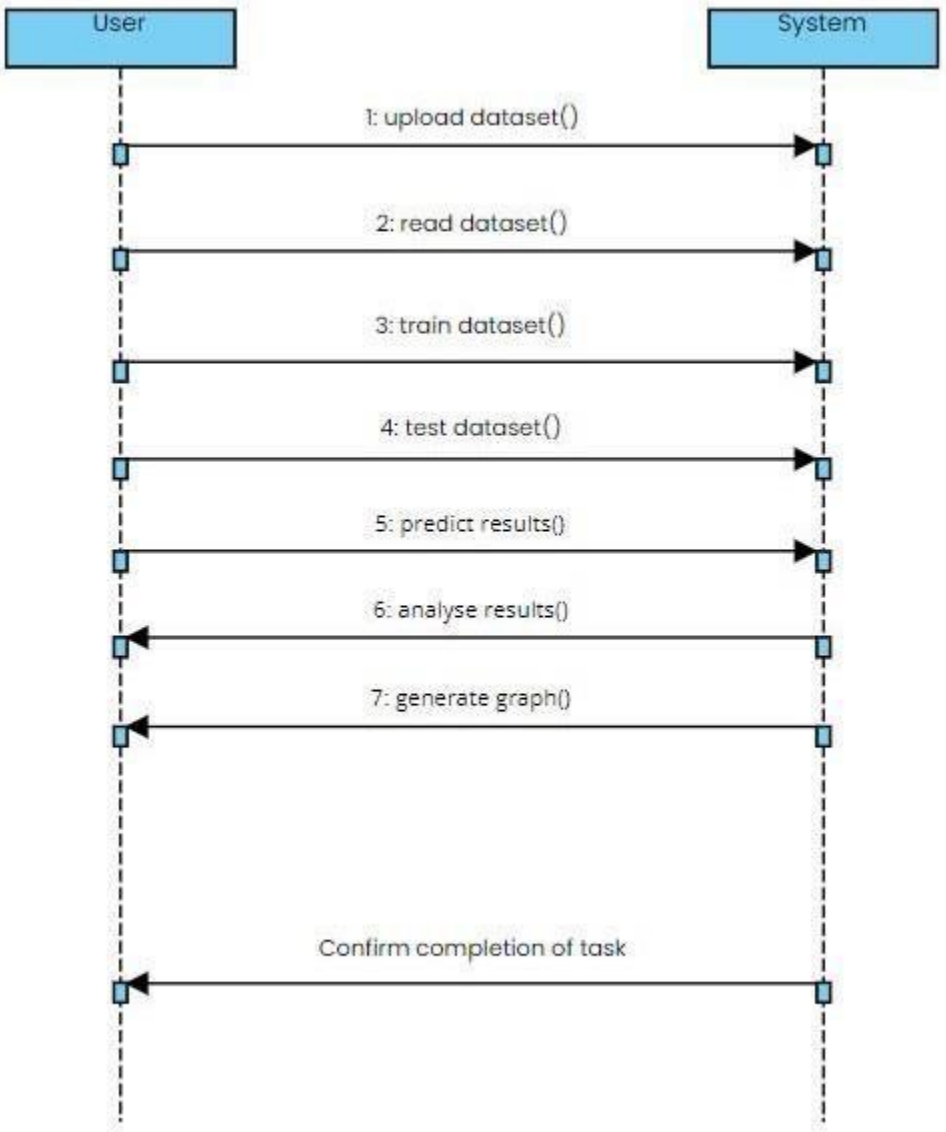

Fig 4.3 Sequence Diagram

4.2.3 CLASS DIAGRAM

A Class Diagram is a type of Unified Modeling Language (UML) diagram that provides a static view of a system by illustrating the classes, theirattributes, methods, and the relationships between them. It serves as a foundation for 21 object-oriented modeling and design, showcasing the structure of the system in terms of its classes and their associations. In a Class

Diagram, classes are represented as rectangles, with the class name at the top, attributes in the middle section, and methods at the bottom. Relationships between classes, such as associations, generalizations (inheritance), and aggregations, are depicted using lines connecting the classes.

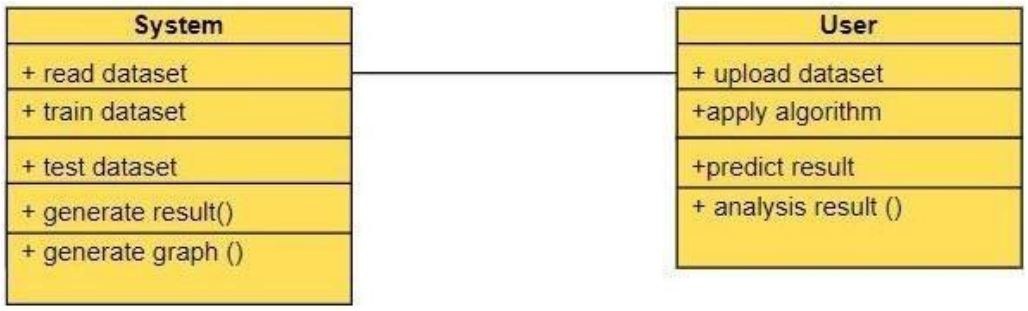

Fig 4.3 Class Diagram

4.2.4 ACTIVITY DIAGRAM

An Activity Diagram is a type of Unified Modeling Language (UML) diagram that depicts the dynamic aspects of a system, focusing on the workflow and flow of activities within a particular process or use case. It provides a visual representation of the sequence of actions, decisions, and parallel activities involved in a specific scenario. In an Activity Diagram, activities are represented by rounded rectangles, and transitions between activities are depicted by arrows. Decision points are shown with diamond shapes, and parallel activities are indicated by branching and merging lines. One of the key advantages of activity diagrams is their ability to provide a clear and intuitive visualization of complex processes, making them an effective tool for communication among stakeholders, including developers, designers, and business analysts.

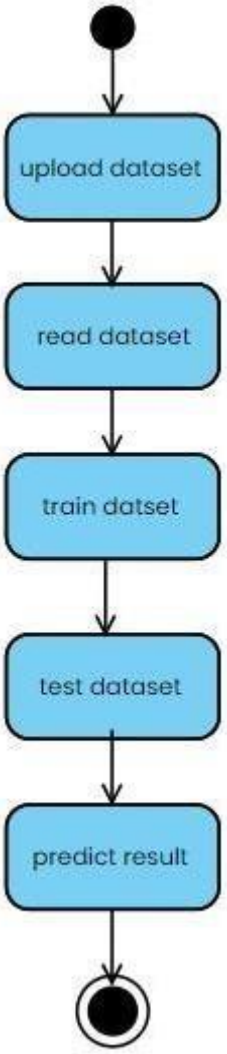

Fig 4.4 Activity Diagram

4.2.5 FLOWCHART DIAGRAM

A flowchart is a visual representation of a process or algorithm, utilizing different shapes and arrows to illustrate the sequence of steps and decision points. It is a diagrammatic tool commonly used in various fields, including software development, business process analysis, and system design, to depict the logical flow of activities. Flowcharts are versatile and effective tools for visualizing complex processes, enabling stakeholders to understand, analyze,

and optimize workflows. They serve as valuable aids in problem-solving, process documentation, and communication among team members and stakeholders.

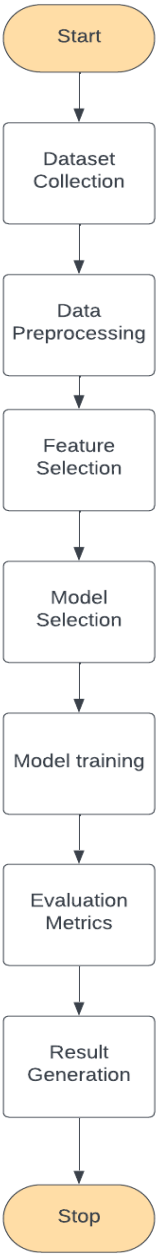

Fig 4.5 Flowchart Diagram

CHAPTER 5

SYSTEM IMPLEMENTATION

5.1 DATA COLLECTION

Data collection for the project "Enhanced Disease Prediction Using Genome-Based Analysis" involves acquiring diverse genomic datasets containing protein IDs as input and disease labels as output. The process begins by identifying reputable sources of genomic data, including public repositories, research databases, and clinical datasets. These sources provide access to a wide range of genomic information, including genetic variants, gene expression profiles, protein-protein interaction networks, and disease phenotypes.

One key aspect of data collection is ensuring the availability and quality of protein ID annotations. Protein IDs serve as the primary input for the disease prediction system, linking genomic data to specific proteins encoded by genes. Therefore, it is essential to obtain protein ID annotations from trusted databases and resources, such as UniProt, NCBI Gene, or Ensembl, which provide accurate and comprehensive information about protein sequences, functions, and associated diseases.

In addition to protein IDs, disease labels are essential for training and evaluating the predictive models. Disease labels indicate the presence or absence of specific diseases or disease phenotypes associated with each protein ID. These labels can be obtained from curated disease databases, clinical registries, electronic health records, or biomedical literature. It is crucial to ensure the consistency and reliability of disease labels by cross- referencing multiple sources and verifying their accuracy through expert curation or validation studies.

5.2 DATA PREPROCESSING

Preprocessing of data for the project titled "Enhanced Disease Prediction Using Genome-Based Analysis," where protein ID serves as input and the output is the name of the disease, involves several critical steps to ensure the accuracy and efficiency of the predictive model. The process encompasses data collection, cleaning, integration, transformation, and feature extraction, each playing a vital role in preparing the dataset for robust analysis and prediction.

Following data cleaning, integration may be necessary if the protein ID information is sourced from multiple databases or datasets. Integration involves merging disparate data sources into a unified dataset, ensuring consistency in format and resolving any conflicts or discrepancies between the datasets. This step facilitates a comprehensive analysis by consolidating relevant information into a single coherent dataset.

Subsequently, data transformation may be employed to prepare the dataset for analysis and modeling. This step involves converting the protein ID information into a suitable format for feature extraction and prediction. Depending on the nature of the data and the requirements of the predictive model, transformation techniques such as encoding categorical variables, scaling numerical features, and dimensionality reduction may be applied.

5.3 FEATURE SELCTION

This involves predicting the name of a disease based on input data consisting of protein IDs derived from genome analysis. To achieve accurate disease prediction, feature selection plays a crucial role in enhancing the performance of the predictive model.

Feature selection is the process of choosing the most relevant and informative features (in this case, protein IDs) from the available data to build a predictive model. In the context of genome-based analysis, there may be thousands or even millions of features (protein IDs) available, but not all of them may be relevant for predicting the occurrence of a specific disease.

To select the most informative features, various techniques can be employed. One common approach is to use statistical methods such as correlation analysis or mutual information to identify the relationship between each feature and the target variable (disease name). Features that have a strong correlation or information content with the target variable are retained, while less informative features are discarded.

Another approach to feature selection is to use machine learning algorithms such as decision trees, random forests, or support vector machines to evaluate the importance of each feature in predicting the target variable. These algorithms can provide a ranking of feature importance, allowing researchers to focus on the most relevant features for disease prediction.

5.4 MODEL SELECTION

This plays a crucial role in determining the effectiveness and accuracy of the disease prediction system. Given that the input to the system is a protein ID and the output is the name of the disease, selecting an appropriate model involves considering several factors such as the complexity of the data, the size of the dataset, computational resources, and the desired level of prediction accuracy.

The choice of model should also consider the interpretability of the results, as it is essential to understand the underlying biological mechanisms driving disease prediction. Additionally, techniques such as cross-validation and

hyperparameter tuning should be employed to optimize model performance and ensure robustness.

Ultimately, the selection of the most suitable model for the "Enhanced Disease Prediction Using Genome Based Analysis" project will depend on the specific characteristics of the dataset, the computational resources available, and the desired balance between prediction accuracy and interpretability. Experimentation with different algorithms and evaluation metrics will be necessary to identify the most effective approach for accurately predicting diseases based on protein ID inputs. This project involves 3 main algorithms namely, KNN, Decision Tree and Naive Bayes.

5.5 MODEL TRAINING

With the dataset prepared and the model chosen, the training process begins. This involves feeding the preprocessed data into the selected model and iteratively optimizing its parameters to minimize prediction errors. During training, the dataset is typically split into training, validation, and test sets to assess the model's performance accurately. Techniques such as cross-validation may be employed to ensure the robustness of the trained model and prevent overfitting.

This process involves hyperparameter tuning wherein Hyperparameters are configuration settings that govern the learning process of the model. Fine-tuning these parameters is essential for maximizing performance. Techniques like grid search or random search can be used to systematically explore different combinations of hyperparameters and identify the optimal settings that yield the best results on the validation set.

5.6 EVALUATION METRICS

"Evaluation metrics" are crucial components in assessing the effectiveness and reliability of a project like "Enhanced Disease Prediction Using Genome-Based Analysis," where protein IDs serve as input for predicting disease outcomes. Given the complexity of this task, a comprehensive evaluation strategy is necessary to gauge the accuracy, efficiency, and overall performance of the predictive model.

One fundamental metric for evaluating the predictive performance of such a model is accuracy. Accuracy measures the proportion of correctly predicted disease outcomes over the total number of predictions made. In the context of disease prediction using genome-based analysis, accuracy reflects the model's ability to correctly identify the disease associated with a given protein ID. High accuracy indicates that the model is making correct predictions, while low accuracy suggests inaccuracies in the predictions, which could lead to misdiagnoses and subsequent treatment errors.

Precision and recall are additional metrics that provide insights into themodel's performance, particularly in scenarios where class imbalances exist among the predicted disease categories. Precision measures the proportion of true positive predictions out of all positive predictions made by the model. It helps evaluate the model's ability to avoid false positives, which are cases where the model incorrectly predicts a disease when it is not present. On the other hand, recall measures the proportion of true positive predictions out of all actual positive instances in the dataset. It assesses the model's ability to capture all relevant instances of a disease, minimizing false negatives where the model fails to predict the presence of a disease when it is actually present.

CHAPTER 6

RESULTS AND EVALUATION

6.1 RESULTS EVALUATION

The performance of three distinct algorithms, namely KNN, Naïve Bayes, and Decision Tree, was assessed for their efficacy in predicting hereditary genomic diseases based on protein IDs, within the framework of the Django Cloud Platform. Leveraging Django's robust backend infrastructure, which facilitated seamless data management and processing, the evaluation was conducted efficiently. KNN demonstrated exceptional predictive capability with an accuracy of approximately 99%, showcasing its effectiveness in classifying diseases by leveraging similarities in genomic data. Similarly, Naïve Bayes demonstrated notable accuracy, achieving approximately 98% accuracy by employing a database-centric approach to match protein IDs with disease names. Contrastingly, the Decision Tree algorithm yielded a lower accuracy of approximately 56%, potentially due to its struggle with the complexity of genomic data. Overall, the analysis underscores the robustness of KNN and Naïve Bayes algorithms in disease prediction, highlighting their potential for enhancing personalized medicine approaches within the Django environment, while suggesting avenues for further optimization of the Decision Tree algorithm's performance in genomic disease prediction workflows.

6.2 PERFORMANCE ANALYSIS

Performance analysis refers to the process of evaluating and assessing the efficiency, effectiveness, and overall success of a system, process, or entity. This analysis aims to measure, understand, and optimize various aspects to ensure optimal functioning.

This project utilized three key algorithms – KNN, Decision Tree and Naïve Bayes. The results of accuracy varied for each algorithm. KNN obtained an accuracy of 99%, Naïve Bayes acquired an accuracy of 98% and Decision Tree obtained an accuracy of 56%. The below pie chart demonstrates how each algorithm varies with different results of accuracy.

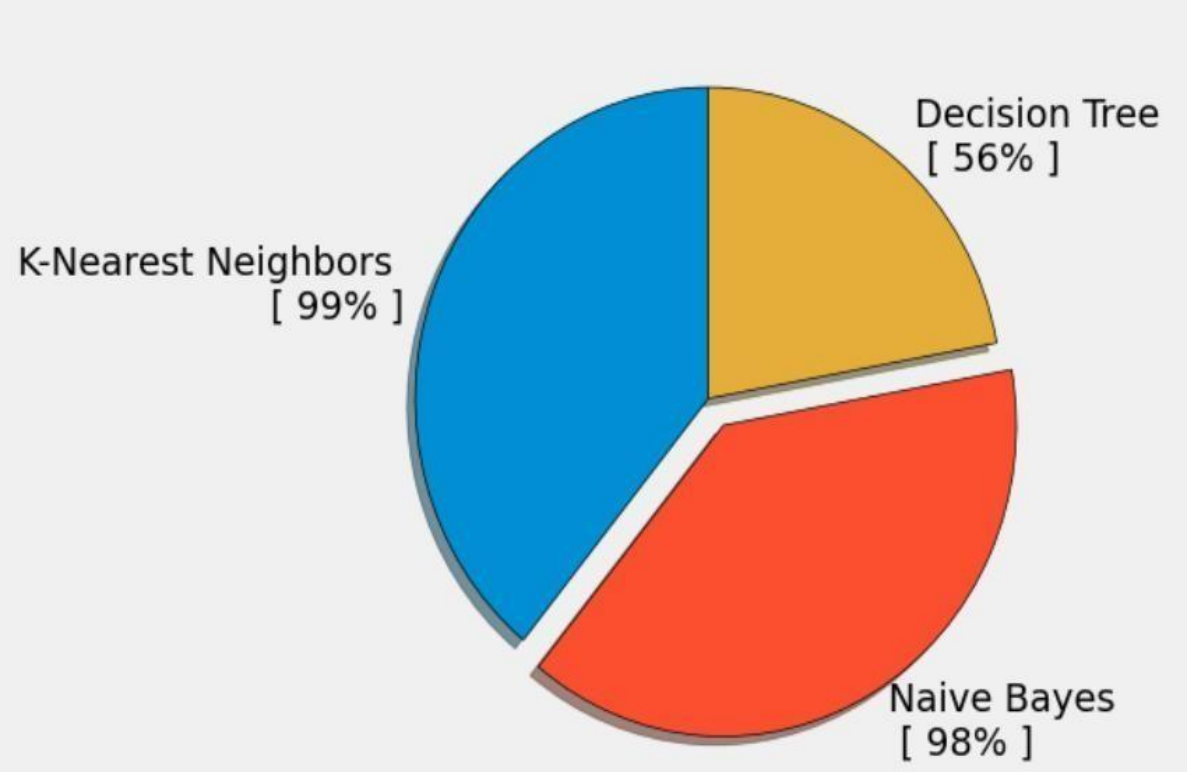

Fig 4.6 Pie Chart Demonstration of Results Comparison

CHAPTER 7

CONCLUSION AND FUTURE ENHANCEMENTS

7.1 CONCLUSION

In conclusion, the project on "Enhanced Disease Prediction Using Genome Based Analysis" marks a significant stride in the realm of medical science and bioinformatics. By leveraging the vast pool of genetic data and employing sophisticated analytical techniques, the endeavor aimed to revolutionize disease prediction and diagnosis. Through the utilization of protein IDs as inputs and the subsequent derivation of disease names as outputs, the project underscores the potential of genomic analysis in unraveling complex relationships between genetic variations and disease susceptibility.

One of the notable achievements of this project lies in its ability to harness the power of genomic data to enhance disease prediction accuracy. By delving into the intricate interplay between proteins and diseases encoded within the genome, the approach offers a more comprehensive understanding of the molecular mechanisms underlying various pathological conditions. This not only facilitates early detection and diagnosis but also paves the way for personalized medicine tailored to an individual's genetic makeup.

Moreover, the project's emphasis on enhanced disease prediction signifies a departure from traditional diagnostic methodologies towards more proactive and preventive healthcare practices. By identifying potential disease predispositions at a molecular level, healthcare providers can intervene preemptively, offering targeted interventions and lifestyle modifications to mitigate the risk of disease development. This proactive approach holds the promise of not only improving patient outcomes but also alleviating the burden on healthcare systems by reducing the incidence of preventable diseases.

7.2 LIMITATIONS

One of the primary limitations associated with the project titled "Enhanced Disease Prediction Using Genome Based Analysis" lies in the complexity and variability of genetic factors contributing to disease manifestation. While utilizing protein IDs as input for disease prediction may offer insights into potential health risks, it's crucial to acknowledge that diseases often result from multifactorial interactions involving numerous genes, environmental factors,lifestyle choices, and epigenetic modifications. Relying solely on protein IDs might overlook crucial genetic variations or mutations that could significantly impact disease susceptibility or progression.

Moreover, the predictive accuracy of the model may be constrained by the availability and quality of genomic data. The completeness and accuracy of genomic databases can vary, leading to potential biases or inaccuracies in disease predictions. Additionally, the model's effectiveness may be limited by the resolution of available genetic information, as certain genetic markers or variations may not be sufficiently captured or annotated in existing databases. This could potentially lead to false negatives or false positives in disease predictions, diminishing the overall reliability of the system.

Another significant limitation pertains to the interpretability and generalizability of the model's predictions. While machine learning algorithms can identify complex patterns and correlations within genomic data, understanding the underlying biological mechanisms driving these predictions remains challenging. The lack of interpretability could hinder the adoption of the model in clinical settings, where transparent and interpretable decision-making processes are essential for informed patient care.

7.3 FUTURE ENHANCEMENTS

In considering future enhancements for the project several avenues can be explored to further refine and improve the predictive capabilities of the system. One potential enhancement lies in the expansion of the dataset used for training and validation purposes. Incorporating a more diverse and comprehensive collection of protein IDs and corresponding disease outcomes could bolster the model's ability to accurately predict a broader range of diseases.

Furthermore, integrating advanced machine learning techniques such as deep learning algorithms could potentially enhance the model's predictive performance. Deep learning models, such as convolutional neural networks (CNNs) or recurrent neural networks (RNNs), have demonstrated remarkable success in various bioinformatics applications, including genomics analysis. By leveraging these sophisticated algorithms, the model may uncover more intricate patterns and relationships within genomic data, leading to more precise disease predictions.

Another avenue for enhancement involves incorporating multi-omics data integration. Genomic data represent just one piece of the puzzle in understanding the molecular basis of diseases. By integrating additional omics data types such as transcriptomics, epigenomics, metabolomics, and proteomics, the model could gain a more holistic view of the molecular mechanisms underlying various diseases. This integrative approach may unveil novel biomarkers or molecular signatures that can further improve disease prediction accuracy.

APPENDICES

Code for the model

```
import pandas as pd
import numpy as np
import warnings
warnings.filterwarnings("ignore")
from sklearn.preprocessing import LabelEncoder
df=pd.read_csv('data.csv', encoding = "ISO-8859-1")
df
df.drop(0,axis=0,inplace=True)
df.columns
df.shape
df=df.drop(['source_id','target_id','weight','source_desc', 'source'], axis=1)
df.info()
df.isnull().sum()
df['target'].value_counts()
df.nunique()
df['target_desc']=df['target_desc'].str[5:]
df
x = df[['target_desc']]
y = df['target']
from sklearn.model_selection import train_test_split
Xtrain, Xtest, Ytrain, Ytest = train_test_split(x,y,test_size = 0.2,random_state
=2)
from sklearn.neighbors import KNeighborsClassifier
KNN=KNeighborsClassifier()
KNN.fit(Xtrain,Ytrain)
print('KNeighborsClassifier Accuracy:',KNN.score(Xtrain, Ytrain))
```

```
from sklearn.naive_bayes import GaussianNB
NB = GaussianNB()
NB.fit(Xtrain,  Ytrain)
print('Train score:',NB.score(Xtrain, Ytrain))
from sklearn.tree import DecisionTreeClassifier
DecisionTree =
DecisionTreeClassifier(criterion="entropy",random_state=2,max_depth=5)
DecisionTree.fit(Xtrain,Ytrain)
print('Train score:',DecisionTree.score(Xtrain, Ytrain))
def prediction(x):
    x=int(x[5:])
    prediction = NB.predict(np.array([[x]]))
    return prediction[0]
result = prediction('DOID:10747')
result
result = prediction('DOID:1492')
import pickle
# Dump the trained Naive Bayes classifier with Pickle
NB_pkl_filename = 'naive_bayes.pkl'
# Open the file to save as pkl file
NB_Model_pkl = open(NB_pkl_filename, 'wb')
pickle.dump(NB, NB_Model_pkl)
# Close the pickle instances
NB_Model_pkl.close()
```

REFERENCES

[1] P.N.Jeipratha, B.Vasudevan "Optimal Gene Prioritization and Disease Prediction using knowledge based ontology structure", in Biomedical Signal Processing and Control 82 (2023) 104548 at ScienceDirect, Jan 2023, doi: 10.1016/j.bspc.2022.104548.

[2] Atta-Ur-Rahman, Muhammad Umar Nasir, Mohammed Gollapalli, Muhammad Zubair, Muhammad Aamer Saleem, Shahid Mehmood, Muhammad Adnan Khan, Amir Mosavi "Advance Genome Disorder Prediction Model Empowered With Deep Learning", in IEEE Access, vol. 10, pp. 70317-70328, 2022, doi: 10.1109/ACCESS.2022.3186998.

[3] Jia-Hao Song, Cui-Xiang Lin, Hong-Dong Li "An Alzheimer's disease gene prediction method based on ensemble of genomewide association study summary statistics", 2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Las Vegas, NV, USA, 2022, pp. 555-560, doi: 10.1109/BIBM55620.2022.9995296.

[4] Belal Medhat; Ahmed Shawish "FLR: A Revolutionary Alignment-Free Similarity Analysis Methodology for DNASequences", in IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 18, no. 5, pp. 1924-1936, 1 Sept.-Oct. 2021, doi: 10.1109/TCBB.2020.2967385.

[5] Ermal Elbasani; Jeong-Dong Kim "AMR-CNN: Abstract Meaning Representation with Convolution Neural Network for Toxic Content Detection", in Journal of Web Engineering, vol. 21, no. 3, pp. 677 -692, May 2022, doi: 10.13052/jwe1540-9589.2135.

[6] Jan Klosa; Noah Simon; Volkmar Liebscher “A Fitted Sparse-Group Lasso for Genome-Based Evaluations”,in IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 20, no. 1, pp. 30-38, 1 Jan.-Feb. 2023, doi: 10.1109/TCBB.2022.3156805.

[7] Imran Makhdoom; Kadhim Hayawi; Mohammed Kaosar; Sujith Samuel Mathew “D2Gen: A Decentralized Device Genome Based Integrity Verification Mechanism for Collaborative Intrusion Detection Systems”, in IEEE Access, vol. 9, pp. 137260-137280, 2021, doi: 10.1109/ACCESS.2021.3117938.

[8] Hala Ahmed, Hassan Soliman, Mohammed Elmogy “Early Detection of Alzheimer’s Disease Based on Single Nucleotide Polymorphisms (SNPs) Analysis and Machine Learning Techniques”, 2020 International Conference on Data Analytics for Business and Industry: Way Towards a Sustainable Economy (ICDABI), Sakheer, Bahrain, 2020, pp. 1-6, doi: 10.1109/ICDABI51230.2020.9325640.

[9] Manuel C´aceres; Brendan Mumey; Edin Husi´c; Romeo Rizz “Safety in Multi-Assembly via Paths Appearing in All Path Covers of a DAG”, in IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 19, no. 6, pp. 3673-3684, 1 Nov.-Dec. 2022, doi: 10.1109/TCBB.2021.3131203.

[10] Neeraj Singla “Genome Based Identification And Visualizing GC-Content Hotspots”, 2023 International Conference on Advancement in Computation and Computer Technologies (In-CACCT), Gharuan, India, 2023, pp. 653-657, doi: 10.1109/In CACCT57535.2023.10141716.

[11] Yi Huang “Codon Effect on the Entire Genome Based upon Genome-Wide Recoded Escherichia coli” 2021 IEEE 9th International Conference on

Bioinformatics and Computational Biology (ICBCB), Taiyuan, China, 2021, pp. 152-156, doi: 10.1109/ICBCB52223.2021.9459235.

[12] Hyein Seo, Yong-Joon Song, Kiho Cho, Dong-Ho Cho, ”Specificity Analysis of Genome Based on Statistically Identical KWords With Same Base Combination,” in IEEE Open Journal of Engineering in Medicine and Biology, vol. 1, pp. 214-219, 2020, doi: 10.1109/OJEMB.2020.3009055.

[13] Kyung-Seop Shin, Byung-Chang Chung, Woo-Chan Kim, Dong-Ho Cho, ”Fast search of locally repetitive elements based on auto-correlation property in genome,” 13th IEEE International Conference on BioInformatics and BioEngineering, Chania, Greece, 2013, pp. 1-4, doi: 10.1109/BIBE.2013.6701582.

[14] Ana Leon, Oscar Pastor, ”Towards a Shared, Conceptual Model-Based Understanding of Proteins and Their Interactions,” in IEEE Access, vol. 9, pp. 73608-73623, 2021, doi: 10.1109/ACCESS.2021.3080040.

[15] Hao Zhang, Chuanxu Yan, Yewei Xia, Jihong Guan, Shuigeng Zhou, ”Causal Gene Identification Using Non-Linear Regression-Based Independence Tests,” in IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 20, no. 1, pp. 185-195, 1 Jan.-Feb. 2023, doi:10.1109/TCBB.2022.3149864.

[16] Jian-Hong Sun, Shi-Meng Ai, Hong-Jun Luo, Bo Gao, ”Estimation of the Equilibrium GC Content of Human Genome,” 2019 IEEE 7th International Conference on Bioinformatics and Computational Biology (ICBCB), Hangzhou, China, 2019, pp.12-17, doi: 10.1109/ICBCB.2019.8854660.

[17] Chenchen Li, Jin Zhao, Haodi Feng, Daming Zhu, ”TransCoord: Genome-guided Transcripts Assembly by Coordinating Candidate Paths into Two-phased

Linear Programming," 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Houston, TX, USA, 2021, pp. 296 -301, doi: 10.1109/BIBM52615.2021.9669376.

[18] Sangseon Lee, Taeheon Lee, Yung-Kyun Noh, Sun Kim, "Ranked k-Spectrum Kernel for Comparative and Evolutionary Comparison of Exons, Introns, and CpG Islands," in IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 18, no. 3, pp. 1174-1183, 1 May-June 2021, doi:10.1109/TCBB.2019.2938949.

[19] Andre Rodrigues Oliveira, Geraldine Jean, Guillaume Fertin, Klairton Lima Brito, Ulisses Dias, Zanoni Dias, "Sorting Permutations by Intergenic Operations," in IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 18, no. 6, pp. 2080-2093, 1 Nov.-Dec. 2021, doi:10.1109/TCBB.2021.3077418.

[20] Qi Niu, Shao-Liang Peng, Xiang-Li-Lan Zhang, Shuai-Cheng Li Ying Xu, Xiang-Cheng Xie, Yi-Gang Tong, "LysoPhD: predicting functional prophages in bacterial genomes from highthroughput sequencing, 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), San Diego, CA, USA, 2019, pp. 1-5, doi:10.1109/BIBM47256.2019.8983280.

[21] Suganya Chandrababu, Dhundy R Bastola, "Identification of Cross-feeding Metabolism Reveals Oral-Microbiome Modulated Host Behavior," 2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Las Vegas, NV, USA, 2022, pp. 2816-2823, doi: 10.1109/BIBM55620.2022.9995013.

[22] Shuai Yuan, Zhaohui Qin, "Read-mapping using personalized diploid reference genome for RNA sequencing data reduced bias for detecting allele-specific expression," 2012 IEEE International Conference on Bioinformatics and

Biomedicine Workshops, Philadelphia, PA, USA, 2012, pp. 718-724, doi:10.1109/BIBMW.2012.6470225.

[23] Hongling Wang, Alberto Maria Segre, Yungui Huang, Jeffrey R. O'Connell, Veronica J. Vieland, "Fast Computation of Human Genetic Linkage," 2007 IEEE 7th International Symposium on BioInformatics and BioEngineering, Boston, MA, USA, 2007, pp. 857-863, doi: 10.1109/BIBE.2007.4375660.

[24] Hongwei Wu, Fenglou Mao, V. Olman and Ying Xu, "Accurate prediction of orthologous gene groups in microbes," 2005 IEEE Computational Systems Bioinformatics Conference (CSB'05), Stanford, CA, USA, 2005, pp. 73 -79, doi:10.1109/CSB.2005.10.

[25] Jian Li, Khalid Sayood, "A genome signature based on Markov modeling," 2005 IEEE International Conference on Electro Information Technology, Lincoln, NE, USA, 2005, pp. 6 pp.-6, doi: 10.1109/EIT.2005.1627006.

[26] Yu Zhou, Li-Qian Zhou, Zu-Guo Yu, Vo Anh, "Distinguish Coding And Noncoding Sequences In A Complete Genome Using Fourier Transform," Third International Conference on Natural Computation (ICNC 2007), Haikou, China, 2007, pp. 295-299, doi: 10.1109/ICNC.2007.333.

[27] Iman Tavassoly, Omid Tavassoly, Mohammad Soltany Rezaee Rad, Negar Mottagi Dastjerdi, "Three Dimensional Chaos Game Representation of Genomic Sequences," 2007 Frontiers in the Convergence of Bioscience and Information Technologies, Jeju, Korea (South), 2007, pp. 219-223, doi:10.1109/FBIT.2007.13.

[28] Jian Li, K. Sayood, "A Genome Signature Based on Markov Modeling," 2005 IEEE Engineering in Medicine and Biology 27th Annual Conference, Shanghai, China, 2005, pp. 2832-2835, doi: 10.1109/IEMBS.2005.1617063.

[29] Jihua Feng, Xianhua Dai, Qian Xing, Zhiming Dai, Jiang Wang, Yangyang Deng, Caisheng He ”A position-slots model for nucleosome assembly in the yeast genome based on integrated multi-platform positioning dataseis,” 2009 Fourth International on Conference on Bio-Inspired Computing, Beijing, China, 2009, pp. 1-6, doi: 10.1109/BICTA.2009.5338121.

[30] H. Terazono, A. hattori, H. Takei, K. Takeda and K. Yasuda, ”Rapid real-time PCR-based nucleotide sequence measurement method using 1480nm infrared laser heating,” 2007 Digest of papers Microprocesses and Nanotechnology, Kyoto, Japan, 2007, pp. 326-327, doi: 10.1109/IMNC.2007.4456236.