“Skin Disease Risk Prediction”

Submitted in partial fulfillment of the requirements of the

University of Mumbai

For the Degree of

Bachelor of Engineering Sem VII of Computer Engineering

Submitted By

Dattaraya Mundhe(28)

Bipin Pal(31)

Priyanshu Sahu(38)

Under Guidance of

Dr. Shankar M Patil

Department of Computer Engineering

Smt. Indira Gandhi College of Engineering

Affiliated to University of Mumbai

(2023-24)

CERTIFICATE

This is to certify that the Project entitled “Early Kidney Disease Prediction” is a Bonafide work of Dattaraya Mundhe(28), Bipin Pal(31), Priyanshu sahu (38) submitted to the University of Mumbai in partial fulfillment of the requirement for the award of the degree of “Bachelor of Engineering” in “Computer Engineering”.

Dr. Shankar M Patil

Practical Incharge

Dr. K T Patil

Head of Department

Dr. Sunil Chavan

Principal

Mini Project [CSL701] Approval

This Project entitled **"Skin disease risk prediction"** by **Dattaraya Mundhe(28), Bipin pal(31), Priyanshu sahu (38)** is approved for the degree of Bachelor of Engineering in Computer Engineering.

Examiners

1……

Internal Examiner Name & Sign

2……

External Examiner Name & Sign

Date:

Place:

DECLARATION

We declare that this written submission repents our own ideas in our words and where other have been included, we have adequately cited and referenced the original sources. We also declare that we have adhered to all principles of academics honestly and integrity and have not misrepresented or fabricated or falsified any idea/fact/data/source in our submission. We understand that any violation of the above will be cause for disciplinary action by the institute and can also awoke penal action from the sources which have thus not been properly cited or from permission has not been taken when needed.

Dattaraya Mundhe(28)

Bipin Pal(31)

Priyanshu Sahu(38)

ABSTRACT

This abstract discusses the relevance of timely and accurate diagnosis and risk prediction in addressing skin diseases. It introduces the Naive Bayes algorithm as a promising tool for this purpose, highlighting its suitability for certain skin disease prediction tasks. The methodology involves data collection, preprocessing, feature selection, model training, and evaluation, with an emphasis on the need for relevant patient data. The Naive Bayes algorithm, known for its simplicity and efficiency, is used for skin disease risk prediction, although it acknowledges the independence assumption's limitations. Collaboration with domain experts, particularly dermatologists, is crucial to ensure accuracy and dataset quality, especially in more complex cases. In summary, this abstract provides a concise overview of using Naive Bayes for skin disease risk prediction, emphasizing the potential, limitations, and the importance of expert collaboration.

ACKNOWLEDGMENTS

This acknowledgment transcends the reality of formality when we would like to express deep gratitude and respect to all those people behind the screen who guided, inspired and helped me for the project. We would like to thank all our friends, all the teaching and non-teaching staff members of the Computer Department, for all the timely help, ideas and encouragement which helped throughout the project.

LIST OF FIGURES

INDEX

CHAPTER 1
INTRODUCTION

1. Introduction

1.1 Introduction

Skin diseases are a widespread and diverse group of medical conditions that affect individuals across the globe. Ranging from common ailments to rare and severe disorders, these conditions have a significant impact on the overall health and quality of life of those affected. Timely and accurate diagnosis and risk prediction play a pivotal role in guiding healthcare practitioners in the effective treatment and management of these conditions. In recent years, the integration of machine learning algorithms into healthcare has emerged as a promising avenue for risk assessment, diagnosis, and treatment support. One such algorithm, Naive Bayes, has proven to be particularly well-suited for certain skin disease prediction tasks, offering a blend of simplicity and efficiency that aligns with the practical demands of clinical practice.

Skin diseases are a widespread and diverse group of medical conditions that affect individuals across the globe. Ranging from common ailments to rare and severe disorders, these conditions have a significant impact on the overall health and quality of life of those affected. Timely and accurate diagnosis and risk prediction play a pivotal role in guiding healthcare practitioners in the effective treatment and management of these conditions. In recent years, the integration of machine learning algorithms into healthcare has emerged as a promising avenue for risk assessment, diagnosis, and treatment support. One such algorithm, Naive Bayes, has proven to be particularly well-suited for certain skin disease prediction tasks, offering a blend of simplicity and efficiency that aligns with the practical demands of clinical practice.

Ultimately, the primary objective of this endeavor is to harness the power of machine learning, Naive Bayes in particular, to develop a robust and practical tool for dermatologists and healthcare practitioners. This tool will empower them to make more accurate skin disease risk assessments, leading to timely diagnoses and tailored treatment strategies, and thereby enhancing patient care and overall public health.

Working of Naive Bayes:

The Naïve Bayes classifier is a popular supervised machine learning algorithm used for classification tasks such as text classification. It belongs to the family of generative learning algorithms, which means that it models the distribution of inputs for a given class or category. This approach is based on the assumption that the features of the input data are conditionally independent given the class, allowing the algorithm to make predictions quickly and accurately.

Step 1. Convert the data set into a frequency table

In this first step data set is converted into a frequency table

Step 2. Create Likelihood table by finding the probabilities

Create Likelihood table by finding the probabilities.

Step 3. Use Naive Bayesian equation to calculate the posterior probability

Now, use the Naive Bayesian equation to calculate the posterior probability for each class. The class with the highest posterior probability is the outcome of the prediction.

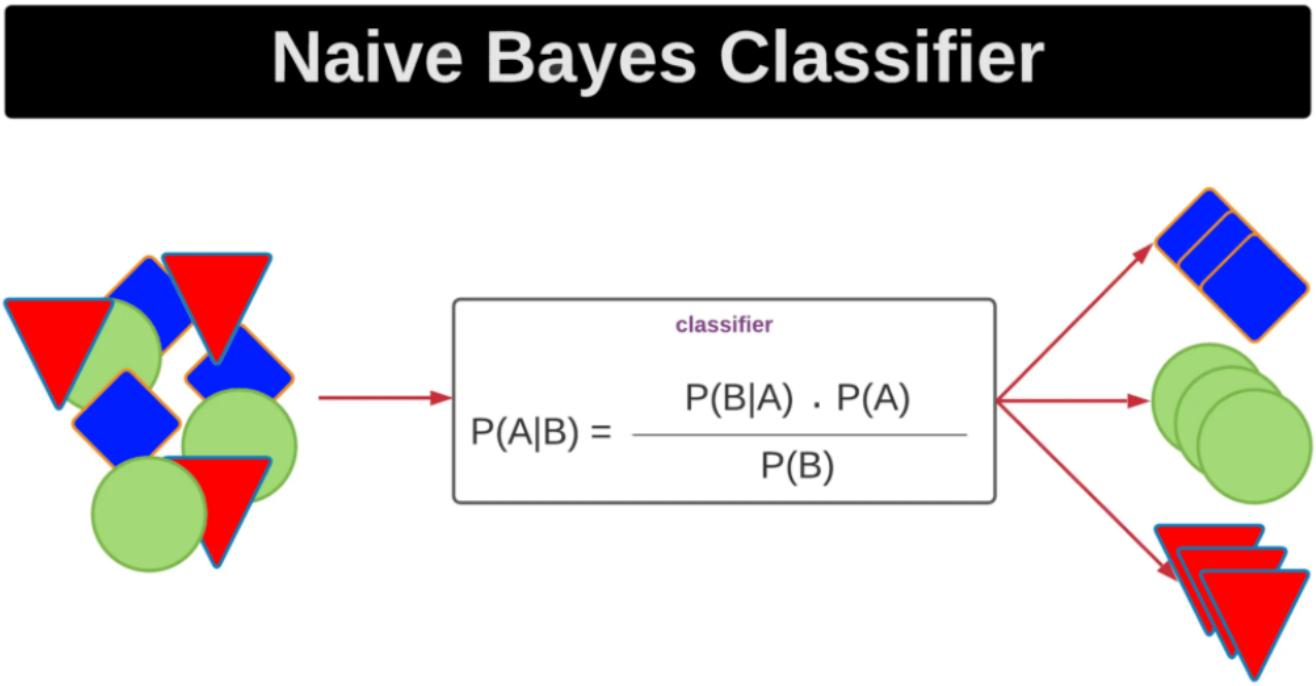

Fig 1.1.1 Working of Naive Bayes

1.2 Problem Statement

In the realm of healthcare, the application of machine learning has become increasingly vital, and skin disease risk prediction is no exception. Timely and accurate diagnosis of skin conditions is essential for effective treatment and patient well-being. To address this, we aim to develop a predictive model using the Naive Bayes algorithm. This model will utilize patient data, including demographics, medical history, genetic predispositions, and various skin-related attributes to assess the risk of skin diseases. By harnessing the power of machine learning, we seek to improve diagnostic accuracy and clinical decision-making. However, the challenge lies in gathering high-quality data, selecting relevant features, training the model, and rigorously evaluating its performance. Moreover, we must address the assumption of feature independence inherent in Naive Bayes, ensuring its suitability for this complex medical task. Collaboration with dermatologists and domain experts is crucial to ensure the model's clinical relevance and accuracy. The successful development of this predictive model has the potential to revolutionize skin disease diagnosis, leading to more timely interventions and improved patient care.

1.3 Objective

- To Gather comprehensive patient information, including demographics and medical history and Handle missing values, encode categorical features, and standardize numerical attributes to make the data suitable for modeling.

- To Determine which patient attributes are most pertinent to skin disease risk prediction and Use appropriate methods to optimize the feature set, promoting model accuracy and efficiency.

- To Utilize the Naive Bayes algorithm for skin disease risk classification and Train the algorithm using a subset of the dataset, adjusting it

- To Use classification metrics like accuracy, precision, recall, F1-score, and AUC-ROC to gauge how well the model predicts skin disease risk and Apply cross-validation techniques to ensure the model's robustness and generalization.

CHAPTER 2
LITERATURE SURVEY

2. Literature Survey

2.1 Literature Survey

Paper 1: Analysis and Classification of Human Skin Diseases

Most common skin diseases like skin cancers, leprosy etc are untreated and mostly causes death. Skin cancer has more cure rate if detected and treated early. The basic means of detecting these skin diseases is through visual inspection followed by biopsy and pathological examination. If the physician finds the appearance of lesion doubtful then normally visual inspection method is used for diagnosis but all malignant lesions are not identified through visual inspection. Now, there are no generally accepted tools that physician can use to immediately find the skin disease in the clinic. Most form of visual inspection could help to prevent misdiagnosis of BCC and other types of skin diseases. Previous work suggests that electrical impedance may distinguish skin cancer from other tissue. The electrical impedance of a tissue depends on its structural characteristics as well as its chemical composition. Studies have shown a wide degree of variation in the bioelectric properties between tissue and cells of body. The studies have shown differences in the electrical impedance of the skin as a result of irritation, allergic reaction, location, sex, age and hydration. A clinical study has also shown significant differences between affected skin and normal skin. Such clinical study is known as impedance measurement and based on a comparison of four indexes: magnitude, phase, real part and imaginary part index.

Paper 2: Skin Disease Classification using Machine Learning Algorithms

Skin diseases are the large number of spread diseases in the world. Their diagnoses are very difficult because of its difficulties in skin texture, presence of hair on skin and color. It is required to develop methods like machine learning in order to increase the accuracy of diagnosis for various types of skin diseases. Machine learning techniques are widely used in medical fields for diagnosis. These algorithms use feature values from images as input to make a decision. The process consists of three stages-The feature extraction stage, the training stage and the testing stage. The process makes use of machine learning technology to train itself with the various skin images. The objective of this process is to increase accuracy of skin disease detection. Three important features in image classification are texture, color, shape, and combination of these. In this work, color and texture features are used to classify the skin disease. Normal skin color is different from the skin with disease. Smoothness, coarseness, and regularity is effectively identified using texture features in the images. Hence, these two features are explored to identify skin disease effectively. In this work, entropy, variance and maximum histogram value of Hue-Saturation-Value(HSV) features are used. These features are used to build machine learning algorithms by using Decision Tree(DT) and Support Vector Machine(SVM). At first level, an entropy measure is used to split the tree. At second level, variance is used to get leafs for textures. In color features, maximum histogram value of the HSV measure is used to split the tree. Accuracy is used to test the performance of the proposed algorithm.

Paper 3: Skin Disease Detection based on Machine Learning Techniques

Skin is the human body's exterior integument. Human skin pigmentation varies from person to person, and skin types include dry, oily, and mixed. The human skin's diversity offers bacteria and other microbes with a diverse home. Melanocytes in the human skin create melanin, which can absorb harmful UV radiation from the sun, causing skin damage and cancer. In most third-world societies, the requisite technologies for early identification of many diseases are still unavailable. The technique of image segmentation aids in the diagnosis of various skin disorders. The goal of this research is to use image processing techniques to diagnose the skin illness from a given image set. Deblurring and noise reduction were performed on the captured image set before it was processed. If acne, dermatomyositis, candidiasis, cellulitis, Scleroderma, chickenpox, ringworm, eczema, psoriasis, Melanoma, and other skin illnesses are left untreated in their early stages, they can lead to a variety of health consequences and even death. The technique of image segmentation aids in the diagnosis of various kind is orders. They have taken two classes here. To classify the condition, there are two types of skin: normal and abnormal. Melanoma and Acnephotos will be processed in the system in the abnormal condition

Paper 4:Machine Learning Approaches for Early Hypertension Prediction: A

Comparative Study

This research delves into various machine learning techniques employed for the early prediction of hypertension. The study compares multiple algorithms, including decision trees, neural networks, and support vector machines, using a diverse dataset. The results showcase the effectiveness of these algorithms in hypertension prediction, laying the foundation for further exploration in this field.

Paper 5:Predictive Modeling of Hypertension Using Deep Learning Techniques

This paper focuses on the application of deep learning techniques, specifically neural networks, for hypertension prediction. The research explores the use of deep architectures and evaluates their performance in comparison to traditional machine learning methods. The study provides insights into the effectiveness of deep learning in handling complex patterns within hypertension data, indicating its potential for accurate predictions.

Paper 6: Hypertension Risk Assessment Using Ensemble Machine Learning Models

This research explores the use of ensemble machine learning models for hypertension risk assessment. By combining multiple base learners, such as decision trees and neural networks, the study constructs an ensemble model that capitalizes on the strengths of individual algorithms. The results demonstrate the superior predictive power of ensemble models, highlighting their potential for accurate and robust hypertension risk assessment.

CHAPTER 3
REQUIREMENT SPECIFICATION

3. Requirement Specification

3.1 Hardware and Software Hardware Requirements:

1. CPU:
    - A modern multi-core processor is essential for training machine learning models efficiently.
    - For more complex models and larger datasets, consider using CPUs with higher clock speeds and multiple cores.
2. RAM:
    - Adequate RAM (Random Access Memory) is necessary to store and manipulate large datasets efficiently.
    - The specific amount of RAM required depends on the dataset size and model complexity. For many projects, 16GB or more is recommended.
3. Storage:
    - You'll need sufficient storage space to store datasets, model checkpoints, and related files.
    - SSDs (Solid State Drives) are preferable to traditional HDDs for faster data access.
4. Internet Connection:
    - A stable internet connection is essential for downloading datasets, libraries, and updates.

3.2 Software Requirements:

1. Operating System:
    - Most machine learning libraries and frameworks are compatible with various operating systems (Windows, macOS, Linux).
2. Python:
    - Python is the primary programming language for machine learning. Ensure you have Python installed (preferably Python 3.x) on your system.
3. Integrated Development Environment (IDE):
    - Choose an IDE or code editor for development. Popular options include Jupyter Notebook, VSCode, PyCharm, and Spyder.

4. Machine Learning Libraries:

- Install the necessary Python libraries for machine learning, such as:
  - NumPy and pandas for data manipulation.
  - scikit-learn for machine learning algorithms, including Naive Bayes.
  - Tkinter for GUI implementation.

5. Version Control:

- Consider using version control systems like Git to track changes in your codebase and collaborate with others.

6. Virtual Environments:

- Utilize virtual environments (e.g., Python's virtual env or conda) to manage project-specific dependencies and avoid conflicts.

7. Dependency Management:

- Use a package manager like pip or conda to install and manage Python packages.

CHAPTER 4
IMPLEMENTATION

4. Implementation

4.1 Development Steps:

1. Dataset Creation:

- We got our data from Kaggle, a trusted source for various datasets, including symptoms of various disease.
- To create a dataset, we first created a Microsoft excel sheet and put the parameter and data which was finalize in the previous step and converting it in Comma Separated Values File (.csv) file

2. Data Preprocessing:

- In data preprocessing, you prepare the dataset for machine learning. In this code:
- The 'prognosis' column is converted from categorical values to numerical values 0 ,1,2,3,4,5 using the map function. This conversion is necessary for most machine learning algorithms to work with categorical data.

3. Feature Selection:

- Feature selection involves choosing the most relevant features from your dataset. In this code:
- The features selected for the model are all symptom columns. These features are chosen as inputs to predict the ' prognosis' target variable.

4. Data Splitting:

- Data splitting is done to create separate datasets for training and testing to assess model performance. In this code:
- The train_test_split function is used to split the dataset into training and testing sets, with a test size of 20%. The random_state parameter is set for reproducibility.

5. Finalizing Algorithm:

- Model selection is the process of choosing an appropriate machine learning algorithm for your task.
- The Naïve Bayes classifier is a popular supervised machine learning algorithm used for classification tasks such as text classification.This approach is based on the assumption

that the features of the input data are conditionally independent given the class, allowing the algorithm to make predictions quickly and accurately.

6. Model Training:

- Model training involves fitting the selected models to the training data.
- The Naive Bayes Classifier are trained on the training data using the fit method

7. Testing the Model

- We used a 20% dataset for testing to see how well the model could spot cataracts.
- Here we also use above step's details for training and testing.
- We find the accuracy for each part and we generate the classification report for each part such as (precision, accuracy, f1-score)

8. Model Evaluation:

Model evaluation is the assessment of how well your model performs.

Accuracy:

- Accuracy is a measure of the overall correctness of the model's predictions. It is calculated as the ratio of correctly predicted instances to the total number of instances in the test set. In the code, accuracy is calculated using the accuracy_score function from the scikit-learn library. A higher accuracy indicates that the model is making more correct predictions.

Precision:

- Precision is a metric that measures the accuracy of the positive predictions made by the model. It answers the question: "Of all the instances predicted as positive, how many are actually positive?"
- In the code, precision is calculated using the precision_score function with the pos_label parameter set to 'Yes'. Precision is particularly important when false positives (incorrectly predicting positive) are costly.

F1 Score:

- The F1 score is the harmonic mean of precision and recall. It provides a balance between precision and recall, particularly useful when there is an imbalance between the number of positive and negative instances in the dataset.
- The F1 score is calculated as 2 * (precision * recall) / (precision + recall). It ranges between 0 (worst) and 1 (best).
- In the code, the F1 score is calculated using the f1_score function with the pos_label parameter set to 'Yes'.

9. Prediction:

- It asks for trained Naive Bayes model to tell you whether a new person, described by the data in input_data, is likely to have Skin disease or not. The model gives an answer, which is stored in the variable predictions.
- Naive Bayes model calculate probabilities for each target class and the one having the highest probabilities will be predicted as output.

4.2 Result Analysis:

Input					Precision	Accuracy	F1 Score
S1	S2	S3	S4	S5	0.91	1.0	1.0

Fig 4.2.1 Skin Disease Risk Prediction 1

Input					Precision	Accuracy	F1 Score
S1	S2	S3	S4	S5	0.91	1.0	1.0

Fig 4.2.2 Skin Disease Risk Prediction 2

	Accuracy	Precision	F1- Score
Fungal Infection	1.0	0.91	1.0
Allergy	1.0	0.91	1.0
Chicken Pox	1.0	0.91	1.0
Acne	1.0	0.91	1.0
Psoriasis	1.0	0.90	1.0
Impetigo	1.0	0.92	1.0
Weighted Avg	1.0	0.90	1.0

Fig 4.2.3 Classification Report

4.3 Result Output:

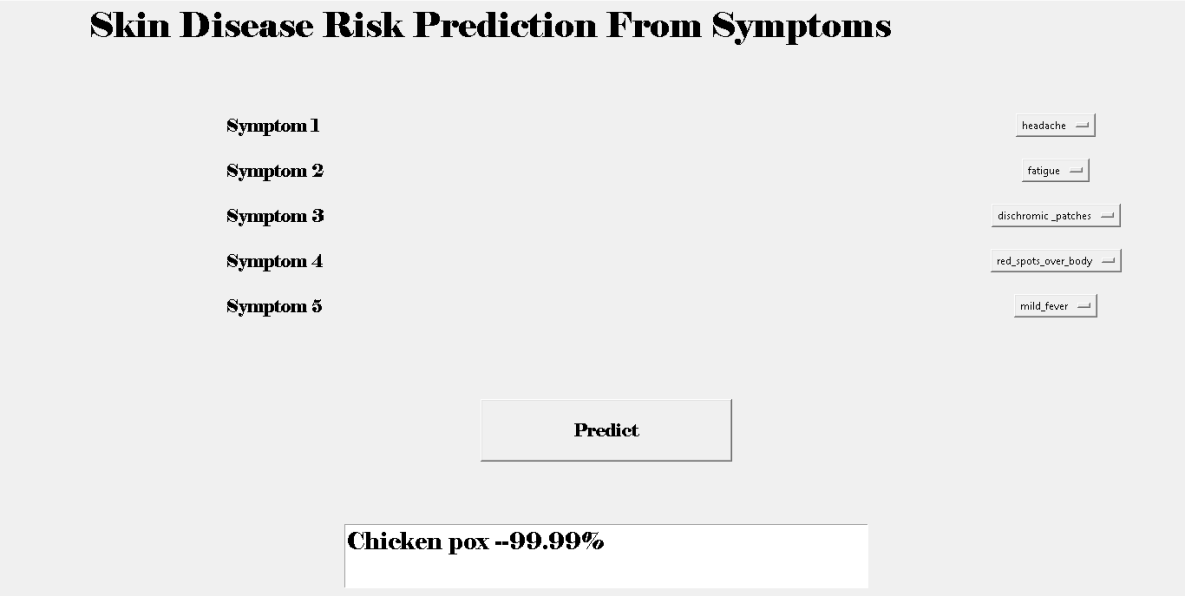

Fig 4.3.1 Output 1

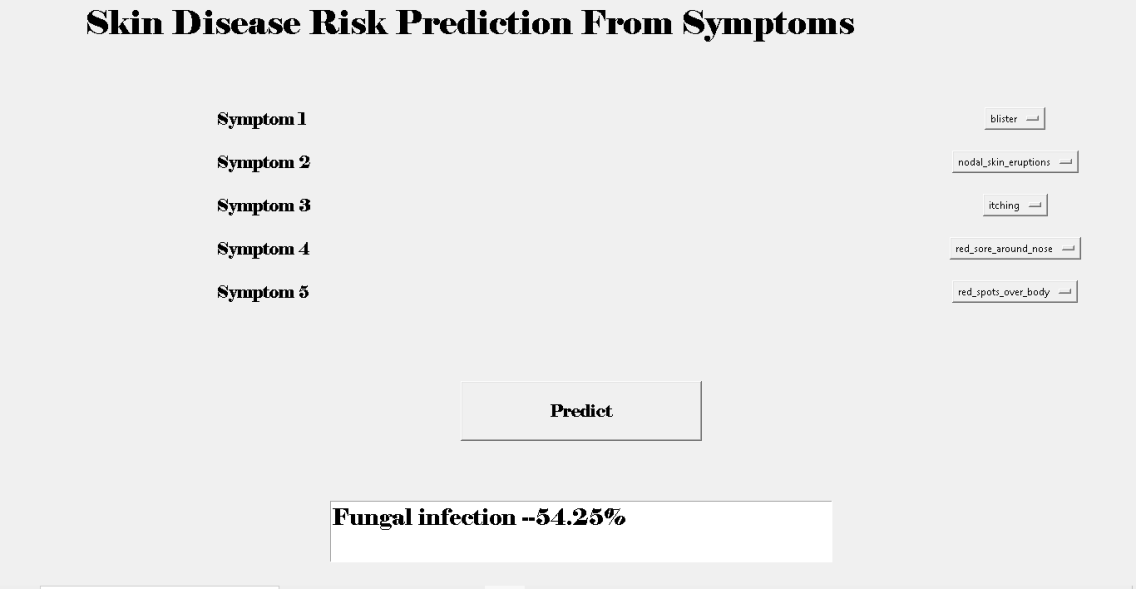

Fig 4.3.2 Output 2

CHAPTER 5
CONCLUSION

5. Conclusion

The utilization of the Naive Bayes algorithm for the development of a skin disease risk prediction model represents a promising approach to enhance dermatological diagnosis and patient care. This journey begins with the critical steps of data collection, cleaning, and ethical considerations. The algorithm's assumption of feature independence holds potential for efficient classification, but it is crucial to address feature interdependencies. Collaborating with domain experts and integrating the model into clinical workflows ensures its clinical relevance and validation. Continuous monitoring, ethical adherence, and addressing challenges are vital for the model's long-term success in improving dermatological healthcare.

CHAPTER 6
FUTURE SCOPE

6. Future Scope

The future scope for skin disease risk prediction using Naive Bayes and other machine learning techniques is promising and multifaceted. Here are some key areas of potential development and advancement:

- Enhanced Accuracy: Continued research and development can lead to improved accuracy by incorporating more comprehensive datasets, advanced feature engineering, and refined modeling techniques. This would make the predictions even more reliable.
- Personalized Medicine: Tailoring predictions to individual characteristics and genetics could offer a more personalized approach to skin disease risk assessment, allowing for precision medicine interventions.
- Telemedicine Integration: Integration with telemedicine platforms can enable remote consultations, with skin disease risk predictions aiding healthcare providers in diagnosing and advising patients from a distance.
- Mobile Applications: Developing user-friendly mobile applications for self-assessment and early detection of skin diseases can empower individuals to take charge of their skin health. These apps can incorporate Naive Bayes models for risk prediction.
- Public Health Initiatives: Expanding the use of these models on a population scale can help public health agencies focus on skin disease prevention efforts, ultimately reducing the burden of skin diseases in communities.

References

1. T. Mitchell, Machine Learning. McGraw-Hill Science, 1997.
2. Neetu Chikyal, K. Veera Swamy, “Performance Assessment of Various Thyroid Image Segmentation Techniques with Consistency Verification”, Vol. 11, 02-Special Issue, 2019,pp.1299-1309.
3. A. Santy and R. Joseph, "Segmentation methods for computer aided melanoma detection," 2015 Global Conference on Communication Technologies (GCCT), 2015, pp. 490-493, doi: 10.1109/GCCT.2015.7342710
4. S. Ahmad, S. Chen, K. Soueidan, I. Batkin, M. Bolic and H. Dajani, “Electrocardiogram assisted blood pressure estimation,” IEEE Transaction Biomedical Engineering , vol. 59, no. 3, pp. 608-618, Mar. 2012.
5. Mas S. Mohktar ,Sami F. Khalil and Fatimah Ibrahim, “The theory and fundamentals of bioimpedance analysis in clinical status monitoring and diagnosis of diseases,” Molecular Diversity Preservation International Open Access Journals, pp. 10895-10928, 2014.