M.Kumarasamy College of Engineering
NAAC Accredited Autonomous Institution
Approved by AICTE & Affiliated to Anna University
ISO 9001:2015 Certified Institution
Thalavapalayam, Karur - 639 113, TAMILNADU.

A Minor Project Report on

DEVELOPING A MACHINE LEARNING MODEL FOR CANCER PROGNOSIS USING RANDOM FOREST TECHNIQUE

Submitted by

RAM PRAKASH S (927622BEE086)

SAPREENA S (927622BEE099)

SRINIDHI B (927622BEE114)

DEPARTMENT OF ELECTRICAL AND ELECTRONICS ENGINEERING

M.KUMARASAMY COLLEGE OF ENGINEERING

(An Autonomous Institution Affiliated to Anna University, Chennai)

THALAVAPALAYAM, KARUR-639113.

MAY 2025

M.KUMARASAMY COLLEGE Of ENGINEERING

(Autonomous Institution, Affiliated to Anna University, Chennai)

BONAFIDE CERTIFICATE

Certified that this Report titled **" DEVELOPING A MACHINE LEARNING MODEL FOR CANCER PROGNOSIS USING RANDOM FOREST TECHNIQUE"** is the Bonafide work of **RAM PRAKASH S (927622BEE086), SAPREENA S (927622BEE099), SRINIDHI B (927622BEE114)** who carried out the work during the academic year (2024-2025) under my supervision. Certified further that to the best of my knowledge the work reported herein does not form part of any other project report.

SIGNATURE	**SIGNATURE**
SUPERVISOR	**HEAD OF THE DEPARTMENT**
Ms.B.Sharmiladevi M.E.,	Dr.J.Uma M.E., Ph.D.,
Assistant Professor	Professor & Head
Department of Electrical and	Department of Electrical and
Electronics Engineering	Electronics Engineering
M.Kumarasamy College of	M.Kumarasamy College of
Engineering, Karur.	Engineering, Karur.

Submitted for Minor Project IV (18EEP401L) viva-voce Examination held at M Kumarasamy College of Engineering, Karur-639113 on ………………..

DECLARATION

We affirm that the Minor Project IV report titled **"DEVELOPING A MACHINE LEARNING MODEL FOR CANCER PROGNOSIS USING RANDOM FOREST TECHNIQUE"** being submitted in partial fulfillment for the award of **Bachelor of Engineering in Electrical and Electronics Engineering** is the original work carried out by us.

REG.NO	STUDENT NAME	SIGNATURE
927622BEE086	**RAM PRAKASH S**	-----------------------
927622BEE099	**SAPREENA S**	-----------------------
927622BEE114	**SRINIDHI B**	-----------------------

VISION AND MISSION OF THE INSTITUTION

VISION

✓ To emerge as a leader among the top institutions in the field of technical education

MISSION

✓ Produce smart technocrats with empirical knowledge who can surmount the global Challenges.

✓ Create a diverse, fully engaged, learner - centric campus environment to provide Quality education to the students.

✓ Maintain mutually beneficial partnerships with our alumni, industry and Professional associations.

DEPARTMENT OF ELECTRICAL AND ELECTRONICS ENGINEERING

VISION

✓ To produce smart and dynamic professionals with profound theoretical and practical knowledge comparable with the best in the field.

MISSION

✓ Produce hi-tech professionals in the field of Electrical and Electronics Engineering by inculcating core knowledge.

✓ Produce highly competent professionals with thrust on research.

✓ Provide personalized training to the students for enriching their skills.

PROGRAMME EDUCATIONAL OBJECTIVES(PEOs)

✓ **PEO1:** Graduates will have flourishing career in the core areas of ElectricalEngineering and allied disciplines.

✓ **PEO2:** Graduates will pursue higher studies and succeed in academic/research careers.

✓ **PEO3:** Graduates will be a successful entrepreneur in creating jobs related to Electrical and Electronics Engineering /allied disciplines.

✓ **PEO4:** Graduates will practice ethics and have habit of continuous learning fortheir success in the chosen career.

PROGRAMME OUTCOMES(POs)

After the successful completion of the B.E. Electrical and Electronics Engineering degree program, the students will be able to:

PO1: Engineering Knowledge: Apply the knowledge of mathematics, science, engineering fundamentals, and an engineering specialization to the solution of complex engineering problems.

PO2: Problem Analysis: Identify, formulate, review research literature, and analyze complex engineering problems reaching substantiated conclusions using first principles of mathematics, natural sciences, and engineering sciences.

PO3: Design/Development of solutions: Design solutions for Complex engineering problems and design system components or processes that meet the specified needs with appropriate consideration for the public health and safety, and the cultural, societal, and environmental considerations.

PO4: Conduct Investigations of complex problems: Use research-based knowledge and research methods including design of experiments, analysis and interpretation of data, and synthesis of the information to provide valid conclusions.

PO5: Modern Tool Usage: Create, select, and apply appropriate techniques, resources, and modern engineering and IT tools including prediction and modeling to complex engineering activities with an understanding of the limitations.

PO6: The Engineer and Society: Apply reasoning in formed by the contextual knowledge to assess societal, health, safety, legal and cultural issues and the consequent responsibilities relevant to professional engineering practice.

PO7: Environment and Sustainability: Understand the impact of the professional engineering solutions in societal and environmental contexts, and demonstrate the knowledge of, and need for sustainable development.

PO8: Ethics: Apply ethical principles and commit to professional ethics and responsibilities and norms of the engineering practice.

PO9: Individual and Teamwork: Function effectively as an individual, and as a member or leader in diverse teams, and in multi-disciplinary settings.

PO10: Communication: Communicate effectively on complex engineering activities with the engineering community and with society at large, such as, being able to comprehend and write effective reports and design documentation, make effective presentations, and give and receive clear instructions.

PO11: Project Management and Finance: Demonstrate knowledge and understanding of the engineering and management principles and apply these to one's own work, as a member and leader in a team, to manage projects and in multi-disciplinary environments.

PO12: Life-long learning: Recognize the need for and have the preparation and ability to engage in independent and life-long learning in the broadest context of technological change.

PROGRAM SPECIFIC OUTCOMES(PSOs)

The following are the Program Specific Outcomes of Engineering Students:

- **PSO1:** Apply the basic concepts of mathematics and science to analyses and design circuits,controls, Electrical machines and drives to solve complex problems.

- **PSO2:** Apply relevant models, resources and emerging tools and techniques to provide solutions to power and energy related issues & challenges.

- **PSO3:** Design, Develop and implement methods and concepts to facilitate solutions for electrical and electronics engineering related real-world problems.

Abstract (Key Words)	**Mapping of POs and PSOs**
Accuracy , Precision , Recall , F1 score , Prostate cancer , Breast cancer, Random forest technique	PO1, PO2, PO3, PO4, PO5, PO6, PO7, PO8, PO9,PO10,PO11, PSO1, PSO2, PSO3.

ACKNOWLEDGEMENT

Our sincere thanks to **Thiru.M.Kumarasamy Founder, and Dr.K.Ramakrishnan B.E, Chairman** of **M.Kumarasamy College of Engineering** for providing extra ordinary infrastructure, which helped us to complete the Minor project in time.

It is a great privilege for us to express our gratitude to our esteemed **Principal Dr.B.S.Murugan., M.Tech., Ph.D.,** for providing us right ambiance for carrying out the project work.

We would like to thank our **Head of the Department Dr.J.Uma.,M.E., Ph.D., Department of Electrical and Electronics Engineering,** for her unwavering moral support throughout the evolution of the project.

We would like to express my deep gratitude to our Minor Project Guide **Ms.B.SharmilaDevi**.,**M.E., Assistant Professor, Department of Electrical and Electronics Engineering,** for his constant encouragement, kind co-operation, valuable suggestions and support rendered in making our project a success.

We offer our wholehearted thanks to our Minor project coordinator **Dr.T.Hariharasudhan.,M.E.,Ph.D., Assistant Professor, Department of Electrical and Electronics Engineering,** for his constant encouragement, kind co-operation and valuable suggestions for making our project a success.

We are glad to thank all the **Faculty Members** of **Department of Electrical and Electronics Engineering** for extending a warm helping hand and valuable suggestions throughout the project.

Words are boundless to thank **Our Parents and Friends** for their constant encouragement to complete this Minor project successfully.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF ABBREVIATION

S.NO	ABBREVIATION	EXPANSION
1	TCGA	The Cancer Gnome Atlas
2	F- SCORE	F1 Score in ML
3	SEER	Seasonal Energy Efficency Ratio
4	ROC	Receiver Operating Characteristics Curve
5	AUC	Area Under the Curve
6	NSCLC	Non Small Cell Lung Cancer
7	SCLC	Small Cell Lung Cancer
8	ML	Machine Learning

/

ABSTRACT

Among the various types of Diseases, Cancer is considered as one of the deadly diseases in the world. Lung, Prostate, and Breast Cancer are some of the Cancer types that are contributing most to the Mortality Rate. In order to overcome, our proposed research work includes Data Collection which is further analyzed and modeled using Machine Learning Techniques. Moreover, the Machine Learning models were evaluated as well as compared based on Performance Metrics parameters like Accuracy, Precision, Recall, F1Score. The random forest Algorithm obtained with the highest Accuracy is Simple Linear Regression for the prediction of Lung Cancer, Logistic Regression for the prediction of Prostate Cancer and Breast Cancer. This has been integrated with python script through the Python Flask Framework. This allows us to fetch the user inputs/responses by the help of Forms and according to the values entered by the users, the interface will predict whether the person is suffering from Benign (Noncancerous) or Malignant (Cancerous). Hence, this study would help in the early detection of Cancer in humans which would give a warning and make them alert regarding the treatment of diagnosed Cancer. It will also help in reducing the Mortality Rate of Cancer patients along with saving of their funds.

Keywords:

Accuracy , Precision , Recall , F1 score , Prostate cancer , Breast cancer, Lung cancer , Random forest algorithm.

CHAPTER 1

INTRODUCTION

Cancer remains one of the leading causes of death globally, and accurate prognosis plays a vital role in determining treatment strategies, assessing survival outcomes, and managing healthcare resources effectively. Prognosis in cancer involves predicting the likely progression of the disease, including survival rates, recurrence probability, and response to therapy. Traditional statistical methods, although valuable, often fall short in capturing the complexity and non linearity inherent in cancer data, which is typically high-dimensional and heterogeneous. With the growing availability of clinical and genomics datasets, machine learning has emerged as a promising approach to improve predictive modeling in oncology. Among various algorithms, Random Forest—a powerful ensemble learning method based on decision trees—has demonstrated significant effectiveness in handling medical prediction tasks. It offers several advantages, such as high accuracy, resistance to overfilling, the ability to manage missing data, and built-in measures for evaluating feature importance. These properties make Random Forest particularly suitable for analyzing large-scale cancer datasets that include a mix of clinical parameters, demographic data, and molecular profiles. Moreover, its ability to model complex interactions between variables allows it to uncover hidden patterns that may be missed by traditional models. Cancer remains one of the leading causes of death worldwide, posing significant challenges to healthcare systems. Early diagnosis and accurate prognosis are crucial for effective treatment planning and improved patient outcomes. In recent years, machine learning (ML**)** has emerged as a powerful tool in the medical field, offering advanced techniques to analyze complex datasets and extract meaningful patterns. In cancer prognosis, ML algorithms can process vast amounts of clinical, pathological, and genomics data to predict disease progression, survival rates, and treatment responses more accurately than traditional statistical methods.

CHAPTER 2

LITERATURE REVIEW

2.1 Comparison of Cancer Morbidity and Mortality

Source:

Cancer Morbidity and Mortality Between Developed Countries, Developing Countries, and China," *2020* ***International Conference*** *on Public Health and Data Science (ICPHDS),* **Guangzhou, China, 2020.**

Inference:

This article is intended to compare the morbidity and mortality of frequent tumour in developed and developing areas, find the popular tumour in China, and compare the results. Data used are those related to global 2018. Lung cancer and breast cancer are the most frequent malignant tumour in developed and developing countries, as indicated by the data analysis. The developed nations have the highest death rates in lung cancer, gastric cancer, liver cancer, and breast cancer.

2.2 Image Based Fractal Analysis for Detection of Cancer Cells

Source:

"Image Based Fractal Analysis for Detection of Cancer Cells," *2020 IEEE International Conference on Bio informatics and Bio-medicine (BIBM),* **Seoul, Korea (South), 2020.**

Inference:

Early diagnosis of cancer is very important for treatment. In this paper, we explored an image based fractal analysis approach for cancer cell detection. Cancer cells normally exhibit abnormalities such as uncontrolled growth of cells. Fractal analysis can employed to examine irregularly shaped objects and the quantify morphological complexity. Image-based experiments using human breast cancer cells were explored.

2.3 Cancer Detection and Analysis Using Machine Learning

Source:

A. Verma, C. K. Shah, V. Kaur, S. Shah and P. Kumar, "Cancer Detection and Analysis Using Machine Learning," *2022 Second International Conference on Computer Science, Engineering and Applications (ICCSEA)***,** Gunupur, India, 2022.

Inference:

In order to overcome, our proposed research work includes Data Collection which is further analyzed and modeled using Machine Learning Techniques. Moreover, the Machine Learning models were evaluated as well as compared based on Performance Metrics parameters like Accuracy, Precision, Recall, F1Score. The best fit Algorithm obtained with the highest Accuracy is Simple Linear Regression for the prediction of Lung Cancer,Logistic Regression for the prediction of Prostate Cancer and Breast Cancer.This has been integrated with python script through the Python Flask Framework.

2.4 Predicting the effectiveness of multi-drug cancer therapies

Source:

Tomic, B. Pirkic, K. Skala and L. Kranjcevic, "Predicting the effectiveness of multi-drug cancer therapies," ***2****019 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, Opatija, Croatia, 2019.

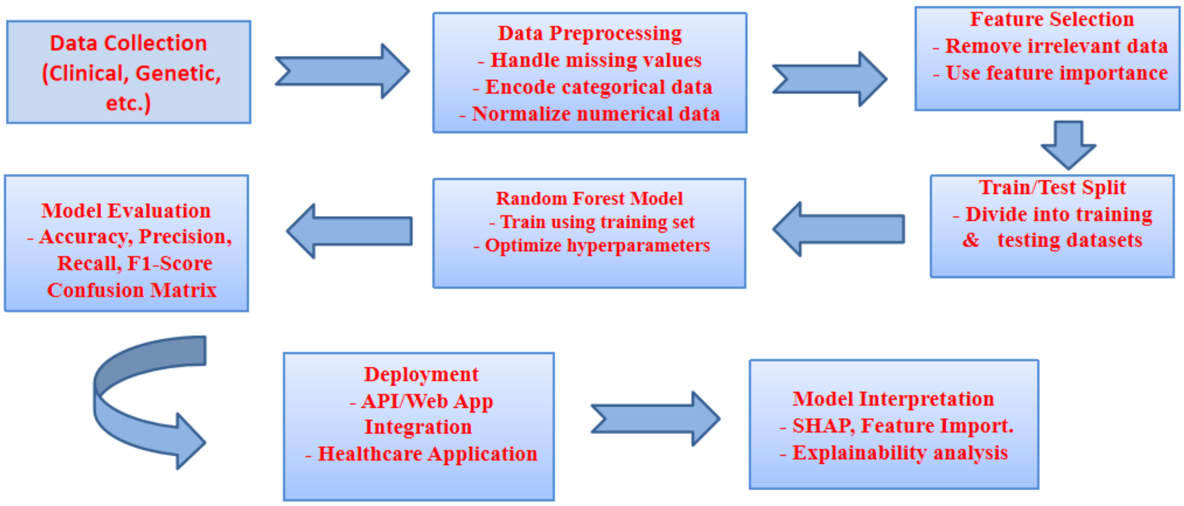
Inference:

The main reason is the ability of cancer to develop resistance against cancer drugs. One strategy to overcome this resistance is to use cancer therapies with several drugs administered at the same time. This can increase our chances to kill cancer cells before they develop resistance. In order to investigate the effectiveness of such therapies, we let the in silicon model of cancer Mini to calculate the most effective 2-drug therapies against non-small cell lung (NSCLC), small cell lung cancer (SCLC), and pancreatic cancer.

CHAPTER 3

PROPOSED METHODOLOGY

3.1 Block diagram



3.2 Fig 3.1 Block diagram of cancer prognosis using random forest technique

Description

This project focuses on building a machine learning model to predict cancer prognosis using the Random Forest algorithm. Cancer prognosis refers to the estimation of a patient's likely clinical outcome, including survival rate, risk of recurrence, and treatment response. Accurate prognosis is crucial in oncology as it influences therapeutic decisions, patient counseling, and healthcare planning. With the increasing availability of large-scale clinical and genomics data, traditional statistical methods often struggle to handle the complexity and dimensional involved. Machine learning, particularly ensemble methods like Random Forest, offers a robust solution for developing predictive models in such data-rich and variable environments comes like overall survival or disease recurrence.

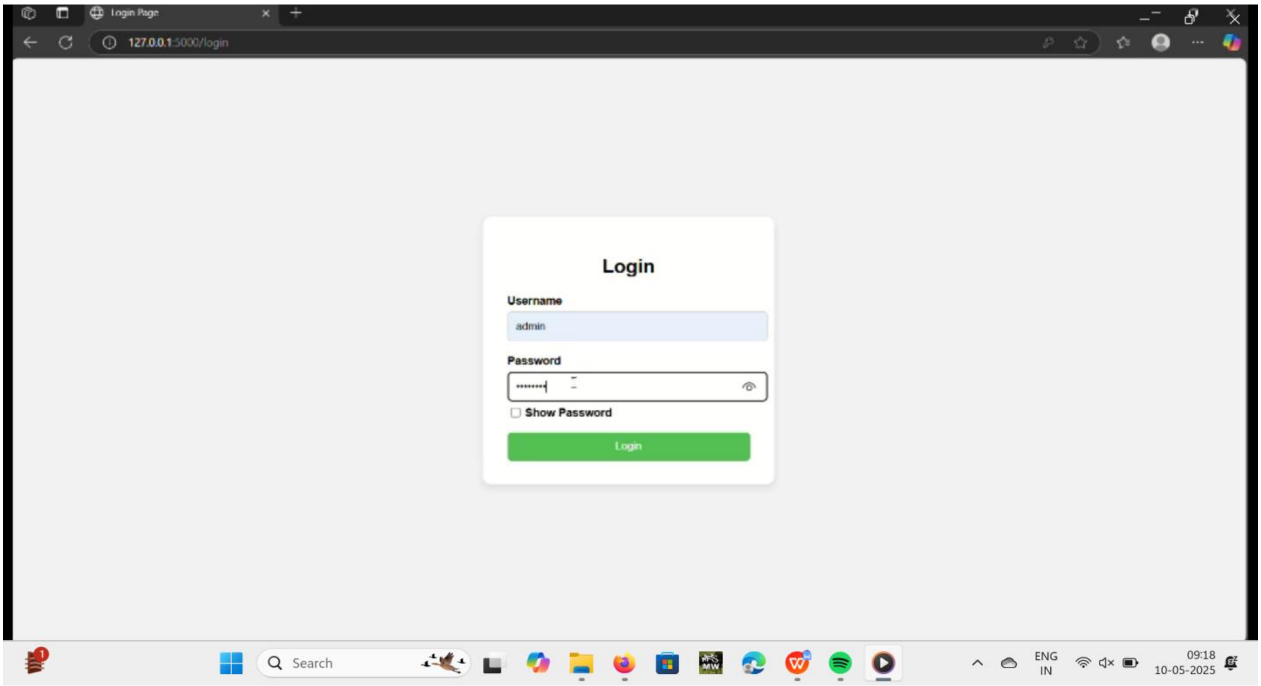
The model development process includes multiple stages: data collection from reliable sources like The Cancer Genome Atlas (TCGA) or SEER, data cleaning and prepossessing (handling missing values, normalizing features, encoding categorical variables), feature selection using statistical and algorithmic methods, model training using Random Forest, and evaluation using metrics such as accuracy, precision, recall, F1-score, and ROC-AUC. Cross-validation techniques are applied to ensure the model generalizes well across different subsets of the data.

One of the key strengths of Random Forest is its ability to provide feature importance scores, which can highlight which variables are most influential in predicting prognosis. This not only improves the interoperability of the model but also offers insights that can support clinical decision-making. The final model aims to serve as a decision-support system that can assist oncologists and researchers in evaluating cancer progression and tailoring treatment plans based on individual patient profiles.

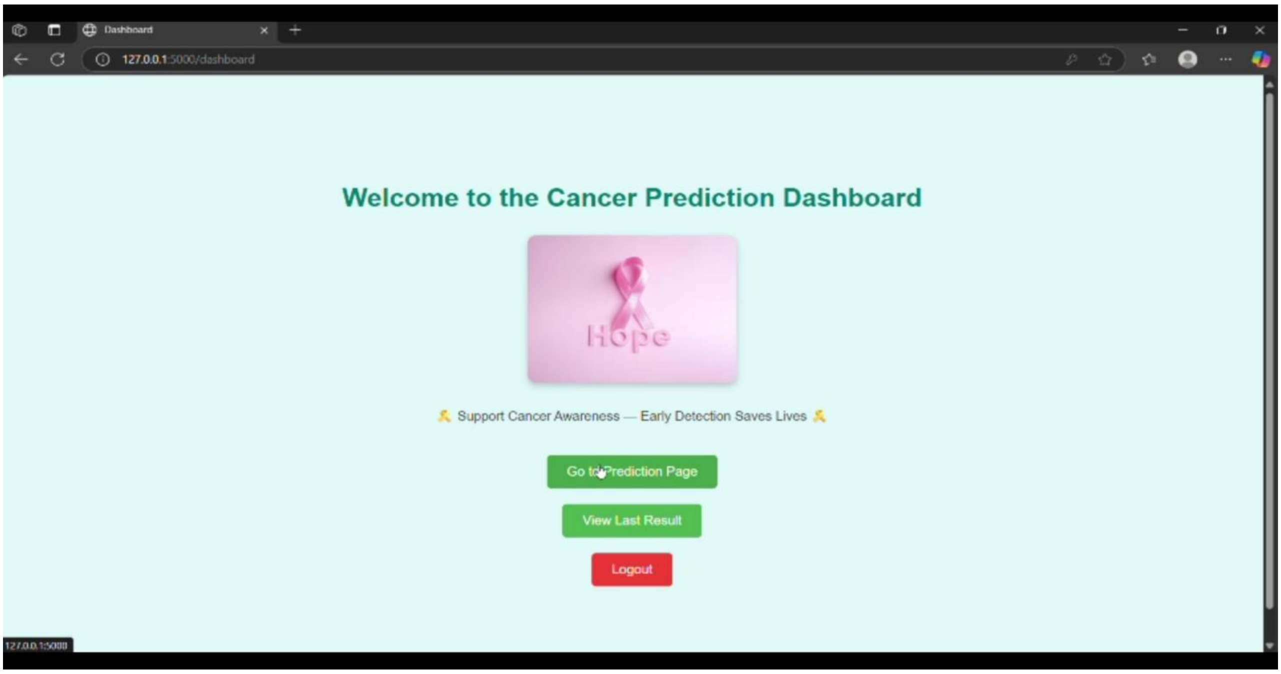
CHAPTER 4

IMPLEMENTATION

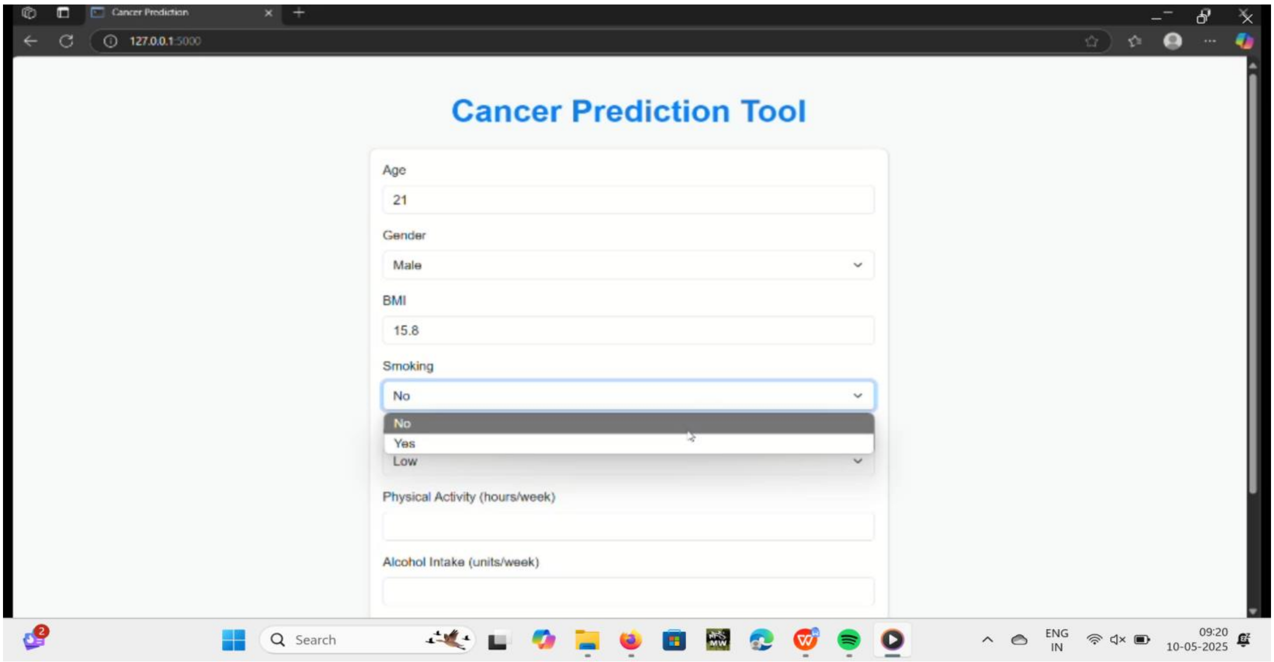
4.1 LOGIN STATUS



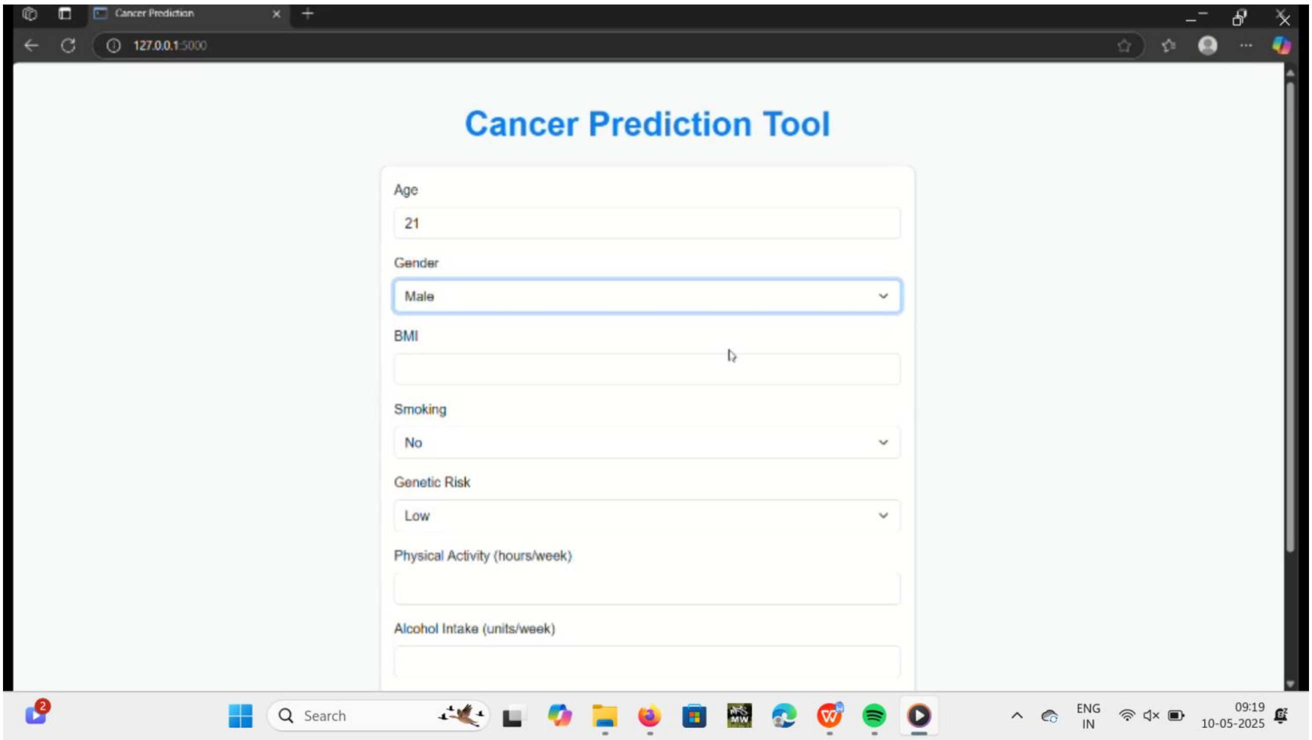
4.2 PREDICTION PAGE



4.3 PHYSICAL ACTIVITY DETAILS



4.4 GENDER DETAILS



CHAPTER 5

RESULTS AND DISCUSSION

5.1 RESULT

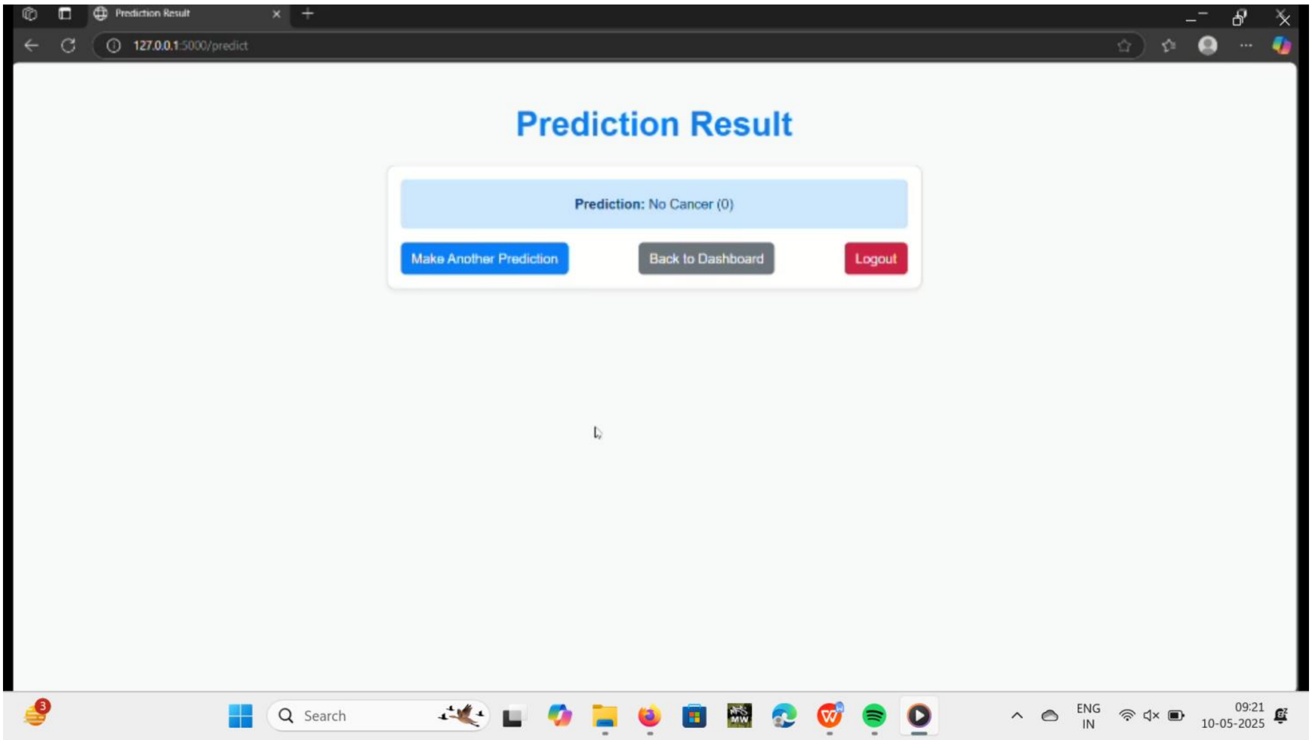


Fig 5.1 Project kit of developing a machine learning model for cancer prognosis using random forest technique

5.2 WORKING EXPLANATION

The project begins with a secure login page designed to ensure that only authorized users, such as healthcare professionals, can access the application. Users are required to enter their username and password, and upon successful authentication, they are directed to the main data entry interface. This next page allows users to input relevant cancer-related patient information, which is critical for accurate prognosis prediction by the machine learning model. The form includes fields such as the patient's age, gender, tumor size, cancer stage, lymph node involvement, and other clinical or pathological features. Once all necessary data is entered, the user submits the form, which sends the information to the back end where the Random Forest model processes it. The model then returns a prognosis prediction, such as the survival probability or a risk classification, which can assist clinicians in making informed decisions about patient care. Once all required data is filled in, the user clicks the submit button, which sends the information to the back-end where the Random Forest algorithm is deployed. The model processes the input data and returns a prediction regarding the patient's cancer prognosis. The output may include survival probability, recurrence risk, or a risk category (e.g., low, medium, high). These predictions are then displayed to the user in a clear and interpretation format, supporting informed clinical decision-making and enabling personalized treatment planning.This data-driven approach enhances decision-making for clinicians and helps personalize treatment plans for patients. By leveraging the capabilities of supervised and unsupervised learning models, researchers can identify key prognostic factors, discover hidden relationships, and ultimately contribute to the development of more effective cancer care strategies.This data is then preprocesed to handle missing values, normalize scales, and remove noise, ensuring that it is suitable for machine learning algorithms.

CHAPTER 6
CONCLUSION AND FUTURE SCOPE

6.1 Conclusion

This project presents a In conclusion, the development of a machine learning model for cancer prognosis using the Random Forest algorithm demonstrates a promising approach to enhancing predictive accuracy and clinical decision-making in oncology. By leveraging large-scale, multi-dimensional datasets that include clinical, demographic, and genomics information, the Random Forest model is capable of identifying complex patterns and interactions that traditional methods may overlook. Its ability to handle missing data, provide feature importance rankings, and maintain high accuracy across varied datasets makes it particularly suitable for medical applications. The project successfully illustrates how such models can be trained and validated to predict critical outcomes such as survival rates and recurrence risks, ultimately contributing to more personalized and effective cancer care. Furthermore, the interoperability of the model allows healthcare professionals to gain insights into the most influential factors affecting prognosis, supporting more informed and data-driven treatment planning. While the results are encouraging, future work may involve integrating additional machine learning algorithms for comparison, incorporating more diverse datasets to improve generalization, and deploying the model in clinical environments for real-time use. Overall, this project highlights the potential of Random Forest as a valuable tool in the ongoing effort to apply artificial intelligence in healthcare and improve patient outcomes in cancer treatment. Despite current challenges such as data quality, model interpretability, and clinical integration, ongoing research and technological advancements continue to strengthen the role of machine learning in oncology. As the field evolves, machine learning will become an increasingly vital tool in the fight against cancer.

6.2 Future scope

1. Expansion to Diverse and Larger Datasets

Future work can involve training the model on larger and more diverse datasets covering various cancer types, stages, and demographics. This will enhance the model's generalization and its applicability across different populations and healthcare settings.

2. Integration of Multi-Comics and Imaging Data

Incorporating additional biological data such as gene expression, proteomics, lamentation profiles, and radiologist images can significantly improve the predictive performance and provide a more holistic view of the disease.

3. Exploration of Advanced and Hybrid Models

While Random Forest is robust, combining it with other machine learning techniques—such as Gradient Boosting Machines (GBM), XGBoost, or deep learning frameworks—can be explored to create hybrid models that improve accuracy and handle complex interactions more effectively.

4. Development of Clinical Decision Support Systems (CDSS)

A practical future goal is to build a user-friendly software or web application that allows clinicians to input patient data and receive real-time prognosis predictions. Integration with hospital electronic health records (EHR) can streamline clinical usage.Efforts toward making machine learning models more explainable and transparent will also improve their acceptance among clinicians, ensuring safer and more trustworthy deployment in clinical settings.Moreover, the integration of real-time data from wearable devices and electronic health records (EHRs) could enable continuous monitoring and dynamic prognosis updates.

5. Emphasis on Explainable AI (XAI)

To improve trust and adoption among medical professionals, future models should include interoperability features. Explainable AI tools like SHAP or LIME can help explain how predictions are made, ensuring transparency and accountability.

6. Clinical Validation and Real-World Deployment

Validating the model through clinical trials or collaborations with hospitals is crucial to assess its real-world reliability. This step is necessary before the model can be confidently deployed in clinical environments.

7. Ethical, Legal, and Regulatory Considerations

Embedding Random Forest models into **real-time clinical decision support systems (CDSS)** for use during patient visits. Potential integration with wearable device data for continuous prognosis monitoring.

8. Explainable AI (XAI) Enhancements

Improving interoperability of Random Forest predictions using tools like SHAP (Shapely Additive Explanations) or LIME. Makes it easier for oncologists to trust and validate the output of the model in clinical practice. Designing adaptive Random Forest models that learn and adjust from new patient data over time.Supports personalized treatment plans tailored to individual risk profiles and responses. Privacy-preserving techniques such as federated learning are also expected to shape the future, enabling collaboration between hospitals and research institutions without compromising patient confidentiality.

REFERENCES

1. Ritu Rani, "Enhanced Lung Cancer Detection using Support Vector Machine Algorithms", *2024 International Conference on IoT Based Control Networks and Intelligent Systems (ICICNIS)*, pp.1583-1587, 2024.

2. Zhen Li, Qijun Yu, Qingyuan Zhu, Xiaojing Yang, Zhaobin Li, Jie Fu, "Applications of machine learning in tumor-associated macrophages", *Frontiers in Immunology*, vol.13, 2022.

3. Drasko Tomic, Jure Murgic, Ana Frobe, Karolj Skala, Antonela Vrljicak, Branka Medved Rogina, Branimir Kolarek, Viktor Bojovic,"Exploring potential therapeutic combinations for castration-sensitive prostate cancer using supercomputers: a proof of concept study", *Scientific Reports*, vol.14, no.1, 2024.

4. Geetha C, Anusha R, Prabhu V, Abinaya Kamatchi S, Gejashree T, "Smart Tool for Earlier Prediction of Breast Cancer Using AI", *2023 International Conference on Research Methodologies in Knowledge Management, Artificial Intelligence and Telecommunication Engineering (RMKMATE)*, pp.1-7, 2023.

5. Prashant A. Athavale, P. S. Puttaswamy, "Gabor Filter Based Ribs Detection on the Chest X-Rays", *2021 5th International Conference on Electrical, Electronics, Communication,Computer Technologies and Optimization Techniques (ICEECCOT)*, pp.596-601, 2021.