Breast Cancer Detection Using Machine Learning

Prerna Sharma1 Prathna Dhall1 Aqdus humza1**Prof. (Dr.) Prerna Chaudhary**1

Department of Computer Science and Engineering,  
Meerut Institute of Engineering & Technology (MIET), Meerut, India  
[prerna.sharma.cse.2021@miet.ac.in](mailto:prerna.sharma.cse.2021@miet.ac.in) ,  
[prathna.dhall.cse.2021@miet.ac.in](mailto:prathna.dhall.cse.2021@miet.ac.in) ,  
[prerna.chaudhary@miet.ac.in](mailto:prerna.chaudhary@miet.ac.in)

**Abstract.** Breast cancer is a serious illness that affects many women around the world. If it is not found early, it can become dangerous and even life-threatening. But if it is detected in the early stages, it becomes easier to treat, and the chances of recovery are much better. Sadly, in many areas, especially in rural or low-income places, people do not have access to quick or easy testing. This delay in checking can lead to late diagnosis, which puts lives at risk.

In this project, we made a smart and simple system using machine learning, a type of technology where we teach a computer to learn from past examples. Just like a student learns from books and pictures, the computer learns from data. We gave the computer real medical data from patients — such as the shape, size, and texture of breast cells — and the computer used this data to understand patterns related to breast cancer.

The system we built can now guess whether a tumor is benign (not harmful) or malignant (harmful) just by looking at a few test values. We used different machine learning models and found that the Random Forest model gave the most accurate results. Then, we created a user-friendly web application where anyone can enter their test values and quickly know the result.

This tool can be very helpful in remote or underdeveloped areas where hospitals and doctors are few. It is fast, easy to use, low-cost, and can help doctors make better and faster decisions. Even though it does not replace a real doctor, it can still help in early detection and guide people to take action quickly. Our goal is to use technology to make healthcare smarter and save more lives by catching breast cancer early.

**Keywords:** Metabolic illness, Neuropathic Ulcer, Blood vessels, Neural Network Models, Foot Ulcer Classification.

1. Inroduction

Breast cancer is one of the most common and serious health problems affecting women globally. It occurs when abnormal cells in the breast grow uncontrollably, forming tumors that can be benign (non-cancerous) or malignant (cancerous). If breast cancer is not detected early, it can spread to other parts of the body, making treatment much harder and reducing the chances of survival. Therefore, detecting breast cancer at an early stage is extremely important because it allows doctors to start treatment sooner, improving the likelihood of successful recovery. Despite the importance of early detection, many people around the world, especially in rural or less developed areas, face difficulties in accessing proper medical tests and diagnostic facilities. Commonly used methods include physical breast examinations, regular doctor checkups, mammograms (X-ray images of the breast), and biopsies (taking tissue samples for testing). However, these traditional methods can have several limitations. They often require expensive equipment, trained specialists, and multiple visits to healthcare centers. In addition, human errors such as misreading test results can sometimes occur, leading to wrong or delayed diagnosis.

In recent years, advances in computer technology and artificial intelligence (AI) have shown great promise in supporting healthcare. One particular branch of AI called machine learning enables computers to learn from data and make predictions without being explicitly programmed for every task. By analyzing large amounts of medical data, machine learning models can recognize complex patterns that may be difficult for humans to detect. In this project, we developed a machine learning-based breast cancer prediction system that helps classify tumors as benign or malignant based on specific features of breast cells. The dataset used for training the model contains measurements related to the size, shape, texture, and other characteristics of cell nuclei from medical tests. These features include radius (which measures how round the cell is), perimeter (the length of the cell’s boundary), area (the size of the cell), texture (variations in the cell surface), and concavity (how curved or indented the edges are). By learning from this data, the model becomes capable of predicting the nature of a tumor when new test results are provided.

Many researchers have previously created similar prediction systems using machine learning. While some models achieved high accuracy, they were often complex and required advanced technical knowledge to use effectively. This made them less accessible to non-experts, busy doctors, or clinics with limited resources. To overcome these challenges, our system focuses on simplicity, speed, and ease of use. We designed a user-friendly web-based interface that allows users to input a few key test values easily and receive a quick prediction. This approach makes the system practical for use in small clinics, community health centers, and even remote areas where advanced medical equipment and specialists might not be available.

It is important to emphasize that this tool is not intended to replace professional medical diagnosis. Instead, it serves as an initial screening assistant that provides an early warning about the possibility of breast cancer. Such early information can encourage individuals to seek timely medical advice, thereby improving health outcomes. This paper is organized as follows: First, we discuss the background of breast cancer and the need for improved early detection methods. Next, we review previous work in this area to understand the strengths and limitations of existing approaches. We then explain the details of our methodology, including the dataset used, the machine learning techniques applied, and the design of the web interface. Following that, we present the results obtained from testing the system and analyze its performance. Finally, we discuss the potential benefits and limitations of our system and suggest directions for future research.

We believe that our machine learning-based prediction system can provide a valuable, cost-effective, and accessible tool that supports early detection of breast cancer, helping to save lives and reduce the burden on healthcare systems.

**2. Literature Review**

Breast cancer is one of the most widespread and dangerous diseases affecting women globally. Its early detection plays a critical role in improving treatment success and increasing survival rates. Traditionally, breast cancer detection has been performed using clinical methods such as physical examinations, mammography, ultrasound imaging, and biopsies. While effective to a degree, these methods often have limitations—they can be expensive, time-consuming, and depend heavily on skilled medical professionals. With the advancement of technology, especially in artificial intelligence and machine learning (ML), researchers have started exploring smarter and more automated ways of diagnosing breast cancer. These modern techniques can help speed up diagnosis, reduce human error, and assist doctors in making more accurate decisions.

2.1 Evolution of Breast Cancer Detection Systems

Earlier, doctors used physical methods such as clinical breast examination (CBE) and mammography, where they would manually check for lumps or use X-rays to detect tumors. These methods were helpful but often failed to detect small or early-stage cancers, especially in younger women with dense breast tissues. To improve diagnosis, researchers started using Computer-Aided Diagnosis (CAD) systems. For example, Madhuri et al. (2016) used image processing techniques to detect abnormalities in mammogram images. While this improved detection, it still required manual analysis by radiologists. Later, researchers began applying machine learning algorithms to analyze patterns in breast cancer datasets. For instance, Patel et al. (2019) developed an ML model using mammogram images and patient data to classify tumors. Their system used algorithms that could learn from past data and achieved higher accuracy than traditional methods. Another example is Saini et al. (2021) who used Convolutional Neural Networks (CNNs) on mammographic images. CNNs automatically learned features from the images, which made the model highly accurate in classifying cancerous and non-cancerous tissues. They reported an accuracy of 92.4% using the MIAS dataset. These research works show how the field has evolved—from manual detection methods to intelligent systems that can predict cancer based on past patterns in data.

****2.2 Common Machine Learning Methodologies****

Most machine learning-based breast cancer detection systems follow a systematic process that begins with data collection. Researchers typically use datasets such as the Wisconsin Breast Cancer Dataset (WBCD), BreakHis, or real-world hospital datasets, which contain various features of breast cell samples, including radius, area, texture, perimeter, compactness, and concavity. These features are essential for distinguishing between benign and malignant tumors. Once the data is collected, it undergoes preprocessing, where missing values are handled, noisy data is removed, and values are normalized or scaled to ensure consistency. This cleaning process helps the model learn more effectively. After preprocessing, feature selection is performed to identify the most relevant attributes that strongly influence the outcome.

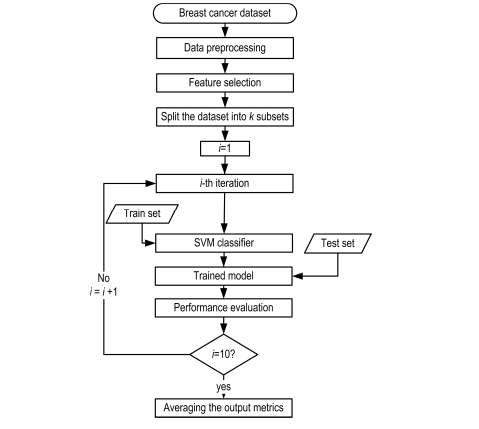
As shown in Fig. 1, the typical workflow for breast cancer detection involves several key steps.

Fig. 1. Typical Workflow of Machine Learning-Based Breast Cancer Detection Systems showing the stages: data collection, preprocessing, feature selection, model training, and evaluation.

For instance, features like radius\_mean and concavity\_worst have been found to be significant indicators of cancer. In one study, Ghosh and Ghosh (2020) applied Principal Component Analysis (PCA) to reduce the number of input features while maintaining model accuracy. Following feature selection, the dataset is used to train various machine learning models. Common algorithms include Logistic Regression (LR), which predicts binary outcomes (cancerous or not); Decision Trees (DT), which split data based on feature values to make decisions; Support Vector Machines (SVM), which find the optimal boundary separating classes; Random Forests (RF), which combine multiple decision trees to improve accuracy; K-Nearest Neighbors (KNN), which classify new cases based on similarity to known cases; and Artificial Neural Networks (ANN), which mimic the human brain to learn complex patterns. For example, Ahmed and Sharma (2020) trained an SVM model using WBCD and achieved over 94% accuracy, particularly using features like radius and concavity. However, they noted that SVM models often struggle when the dataset is imbalanced. After training, the models are evaluated using test data that the model has not seen before. Performance metrics such as accuracy, precision, recall, and F1-score are used to assess how well the model performs. Researchers like Joshi and Mehra (2022) demonstrated that ensemble models like Random Forest and Gradient Boosting often yield more reliable predictions than single models, as they reduce overfitting and combine the strengths of multiple algorithms.

****2.3 Challenges in Existing Research****

Despite the promising results of ML models in breast cancer detection, several challenges persist. One major issue is the lack of real-world and diverse data. Most research relies on publicly available datasets like WBCD, which are relatively small and often lack demographic diversity. As Kapoor and Singh (2023) pointed out, models trained on data from specific regions or age groups may not generalize well to other populations. Another major concern is the use of “black box” models such as deep neural networks and convolutional neural networks (CNNs), which, although highly accurate, do not provide clear reasoning for their predictions. This lack of explainability makes it difficult for medical professionals to trust and adopt these models. Das and Roy (2023) emphasized the importance of Explainable AI (XAI) to bridge this trust gap and make machine learning outputs more transparent and interpretable. Overfitting is another common issue, where a model performs well on training data but poorly on unseen data, indicating it has memorized patterns rather than learning general rules. Additionally, many high-performing models require substantial computational resources, including large memory and GPU power, which may not be available in small clinics or rural hospitals.

****2.4 Gaps Identified in Previous Research****

Several gaps in existing literature have been identified. Most notably, many machine learning models are not tested with real-time clinical data obtained from hospitals, limiting their practical applicability. Furthermore, few studies focus on region-specific datasets, such as data from Indian cities like Meerut, which affects the model’s ability to adapt to local patient demographics and conditions. Another limitation is the lack of explainable models that clinicians can easily understand and trust. In addition, existing studies often rely solely on structured clinical data or imaging data, rather than combining both, which could enhance the prediction accuracy and depth of diagnosis.

****2.5 Suggestions for Future Research****

To address the above challenges, future research should focus on gathering larger, more diverse, and region-specific datasets, particularly from hospitals in underrepresented areas like Meerut. This would allow the development of models that are better suited for local populations. Furthermore, researchers should aim to create lightweight models that can run efficiently on low-cost or resource-constrained devices, making them ideal for deployment in rural clinics. There is also a growing need to adopt explainable algorithms such as decision trees or interpretable neural networks so that medical professionals can understand and validate the predictions. Integrating both medical imaging and clinical test data into hybrid models can lead to more comprehensive and accurate diagnoses. Additionally, implementing real-time predictive systems could support doctors during patient consultations and improve timely diagnosis and treatment.

Table 1. Summary of Key Research Studies on Machine Learning-Based Breast Cancer Detection

| **Author(s) & Year** | **Method / Algorithm Used** | **Dataset** | **Accuracy / Result** | **Key Observations** |
| --- | --- | --- | --- | --- |
| Madhuri et al. (2016) | Image Processing (CAD) | Mammogram Images | Not specified | Improved detection over manual methods |
| Patel et al. (2019) | ML on Image + Patient Data | Mammogram Dataset | Higher than traditional | Used hybrid model for classification |
| Saini et al. (2021) | CNN | MIAS | 92.4% Accuracy | High accuracy using deep learning |
| Ghosh & Ghosh (2020) | PCA + ML Classifiers | WBCD | Not specified | Dimensionality reduction using PCA |
| Ahmed & Sharma (2020) | SVM | WBCD | Over 94% Accuracy | SVM performed well but struggled with imbalanced data |
| Joshi & Mehra (2022) | Ensemble Models (RF, GBM) | WBCD | Higher than single models | Reduced overfitting using ensemble techniques |
| Kapoor & Singh (2023) | Diverse Data Analysis | Regional Data | Not specified | Emphasized lack of generalization |
| Das & Roy (2023) | XAI Models | WBCD | Not specified | Focused on explainability and model trust |

****2.7 How Our Project Builds on This Work****

Our project integrates these valuable insights from existing research. We have carefully selected features such as radius\_mean, concavity\_mean, and area\_worst based on their proven significance in previous studies. To ensure both accuracy and interpretability, we have chosen machine learning algorithms like Random Forest and Logistic Regression, which provide high performance while remaining transparent. A key goal of our project is to design a simple and efficient system that can be used even in resource-limited settings like local hospitals without access to high-end equipment. Additionally, we emphasize a user-friendly interface and explainable results, making it easier for doctors to understand and rely on the model’s predictions during diagnosis.

3. ****Proposed Methodology of Our Work****

Our project aims to develop an intelligent computer-based system that helps in the early detection of breast cancer by analyzing medical data using machine learning. The goal is to provide an accessible and accurate tool that can assist both doctors and patients in identifying whether a breast tumor is likely to be cancerous (malignant) or non-cancerous (benign). This system works by learning from real medical data and identifying patterns that indicate the presence of cancer. The entire process involves several key steps: gathering and preparing the data, training a machine learning model to recognize tumors, making predictions, and presenting the results through a user-friendly interface.

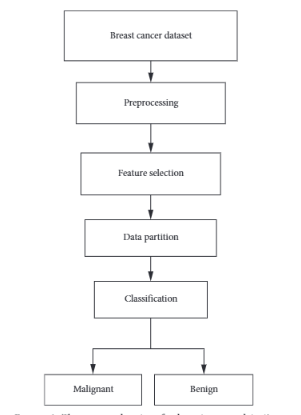


Figure 2. Proposed Breast Cancer Detection System Architecture

The general structure of the system is divided into three main stages: data handling, model development, and user interaction. In the first stage, we collect clinical data from trusted datasets, such as the Breast Cancer Wisconsin dataset. This data contains various measurements taken from breast tissue samples, including tumor radius (size), texture, perimeter, area, smoothness, compactness, concavity, and more. To ensure the system understands the data properly, we clean and normalize the dataset by removing any missing values or inconsistencies. This step also includes scaling the features so that all values are in a similar range. In the second stage, we train the machine learning model using this processed data. The system uses an algorithm called **Random Forest,** which is made up of many decision trees that each analyze different aspects of the tumor data. Together, they make a final prediction by voting on whether the tumor is benign or malignant. This algorithm is chosen because it is known for high accuracy and reliability, especially when working with complex medical data. One important capability of the model is **detecting the tumor and calculating its size** using features such as "radius mean," "area mean," and "perimeter worst."

Table 2. Key Features Used in the Model and Their Medical Significance

| **Feature Name** | **Description** | **Role in Detection** |
| --- | --- | --- |
| Radius Mean | Average radius of tumor | Indicates tumor size |
| Area Mean | Average area covered by the tumor | Larger areas often linked to malignancy |
| Perimeter Worst | Perimeter of the tumor in worst case | Helps assess spread and border irregularity |
| Concavity Mean | Degree of inward curves in tumor edges | High values suggest malignancy |
| Concave Points Worst | Most severe points where edges curve inward | Reflects irregular shape – a cancer marker |
| Texture | Smoothness/roughness of tissue image | May indicate tumor consistency |

These features represent the physical size of the tumor as seen in imaging or tests. The system uses these values to not only detect if a tumor is present but also assess its growth, structure, and potential risk based on its dimensions.

After the model is trained and tested, we provide a simple and informative user interface that allows users to input medical test results. When a doctor or patient enters values like tumor radius or concavity into the input screen, the system processes this data and quickly gives a prediction about the nature of the tumor. The interface also highlights which input values played the biggest role in the decision, making the prediction more transparent. Additionally, the result screen displays a confidence score and uses color coding to help users understand the outcome easily. For those who want to know more, a separate insights screen is available, which gives a deeper explanation of how the system arrived at the conclusion. This includes a visual analysis of which tumor characteristics influenced the prediction the most.

The system itself is divided into different modules, each with a specific role. The **Data Collection Module** is responsible for gathering medical data, especially measurements from tumor images. The **Data Preprocessing Module** cleans this data and prepares it for machine learning. Then, the **Feature Extraction Module** picks out the most important characteristics, such as tumor size, shape, and texture. These features help the system make better predictions. The **Prediction Module** is the brain of the system; it uses the trained Random Forest model to classify the tumor. Finally, the **User Interface Module** presents everything clearly to the end-user, ensuring that the tool can be used easily in hospitals or clinics, or even for research purposes.

By combining all these modules and processes, our breast cancer detection system becomes a reliable and efficient tool that can detect tumors, analyze their size, and predict their nature with confidence. The use of machine learning not only makes the system smarter over time but also supports doctors in making more accurate and faster decisions—especially important in the early diagnosis of breast cancer, which can save lives.

4. Experimental Result and Discussion

In this section, we discuss what we found after building and testing our breast cancer prediction system. We used a smart machine learning model to predict whether a tumor is **benign (non-cancerous)** or **malignant (cancerous)**. To do this, the model studied various features that describe the tumor’s shape, size, and texture. These results were represented not only in numbers but also through visual tools like **bar charts**, **line graphs**, and a **correlation heatmap** to make them easy to understand and interpret.

****4.1 Understanding the Dataset and Tumor Characteristics****

We worked with a well-known dataset called the **Wisconsin Breast Cancer Dataset**, which contains measurements of tumors from past patients. Each tumor is labeled as either **malignant (0)** or **benign (1)**. To make our model faster and more accurate, we selected the **10 most important features** from the dataset. These features included measurements like **radius mean** (average width of the tumor), **area mean** (total space taken up by the tumor), and **concavity mean** (how deeply the tumor curves inward). We also looked at the **“worst”** values for each feature, which represent the most extreme or severe measurements recorded for that tumor. One of the key goals was not only to **detect the tumor** but also to **understand its size and structure** clearly. For example, the **radius mean** tells us how wide the tumor is on average, while **perimeter mean** tells us the outer edge measurement. **Area mean** indicates the surface space covered by the tumor. If these values are high, the tumor is likely to be large. Additionally, features like **concavity mean** and **concave points mean** give us insight into the shape irregularity of the tumor. Malignant tumors often have **sharper, deeper, and more uneven edges**, which are captured by these shape-based features. By comparing benign and malignant tumors, we noticed that **malignant tumors tend to have bigger sizes, more inward curves, and more concave points**. This means that tumors that are larger and more irregular in shape are more likely to be cancerous. Therefore, analyzing both **size and shape** together is crucial for accurate detection.

To better understand how these features relate to one another, we created a pair plot (Figure 3) that displays scatter plots comparing pairs of tumor features such as radius worst, area worst. Each dot represents a patient’s tumor, with orange dots indicating malignant tumors and blue dots benign tumors. This visualization highlights the distinct clusters formed by benign and malignant cases, aiding in visual classification.

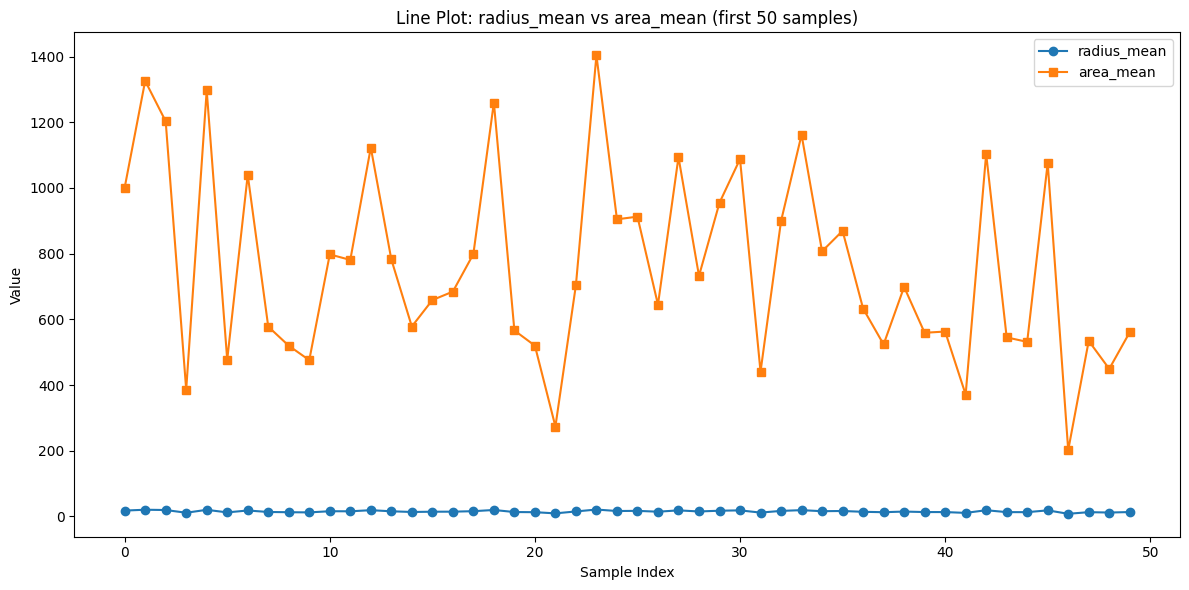
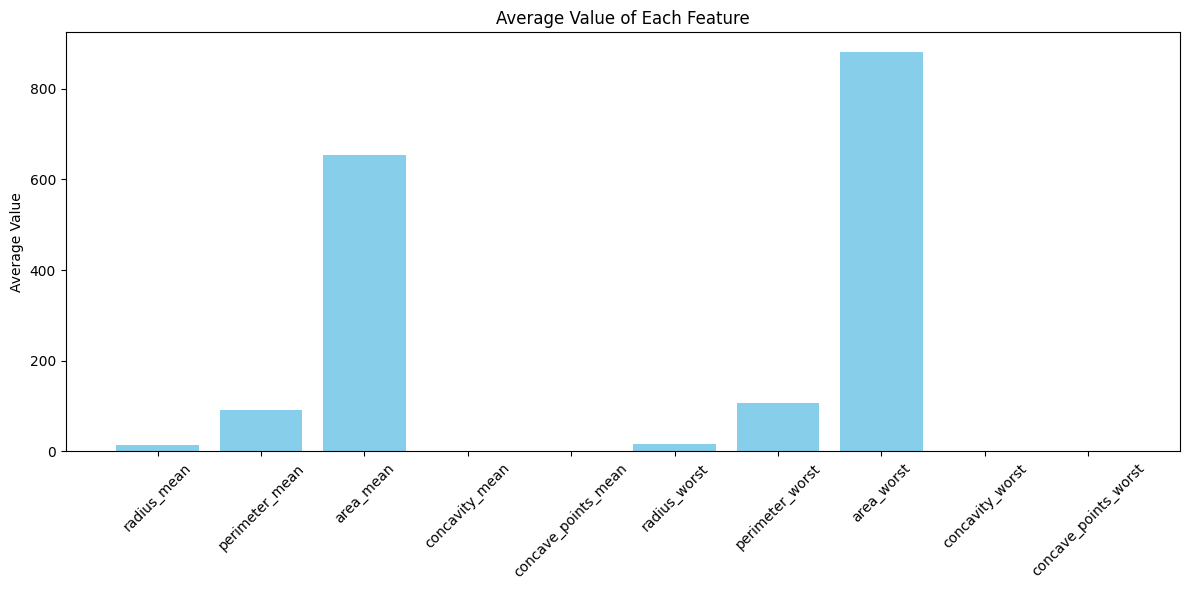


Figure 3: Pair Plot Showing Relationships Between Tumor Features, Colored by Tumor Type (Benign = Blue, Malignant = Orange)](path\_to\_pairplot\_image)

* 1. ****Bar Charts and Line Graphs: Making the Data Easy to See****

To visualize our findings, we created **bar charts** that showed the **average values** of each feature across all tumor samples. Features like **area worst** and **radius mean** had very high average values for malignant tumors—area worst went above 800 (square units), showing that dangerous tumors cover a lot of space. Meanwhile, features like **concavity mean** and **concave points mean** had much smaller average values, often near 0. But this doesn’t make them less important—they still play a big role in identifying sharp, unusual edges that are common in cancerous tumors. We also created **line graphs** that followed the feature values across different tumor samples. These graphs showed that the feature values of malignant tumors often rise quickly, creating sharp upward patterns. This kind of steep increase makes it easier to **spot dangerous tumors early**, even just by visual observation.

Figure 4 displays the average values of selected tumor features for benign and malignant cases. This visual representation highlights how malignant tumors tend to have significantly higher values for features like area worst and radius mean.



**Figure 4: Average Feature Values for Benign vs. Malignant Tumors**

****4.3 Understanding Feature Relationships: The Correlation Heatmap****

To find out how different tumor features are related, we built a **correlation heatmap**, which uses color to show how strongly two features are connected. For example, if two features always increase or decrease together, they show a **strong positive correlation** (close to +1) and appear in **red**. If they go in opposite directions, they show a **negative correlation** (close to -1) and appear in **blue**. One important observation from the heatmap was that **radius mean**, **perimeter mean**, and **area mean** had a **very strong correlation (0.99)** with each other. This means that as the tumor’s radius increases, its perimeter and area also increase—which makes sense because these measurements are related by geometry. Similarly, the “worst” values like **radius worst**, **area worst**, and **perimeter worst** also showed strong correlations. We also saw strong relationships among shape features. For example, **concavity mean and concave points mean** had a high correlation of **0.92**, indicating that tumors with more sharp inward curves also had more concave points. These features are especially useful in detecting **malignant tumors**, which are often more irregular in shape. On the other hand, some features had only moderate or weak correlations. For instance, **concavity worst and radius mean** had a correlation of **0.53**, meaning that even **small tumors can sometimes be very irregular**, and that irregularity is still a warning sign. This highlights the importance of not relying on size alone to determine how dangerous a tumor might be.

Figure 5 shows the correlation heatmap, highlighting how certain features (e.g., radius mean, area mean, and perimeter mean) are highly correlated. Red indicates strong positive correlation, and blue indicates negative correlation.

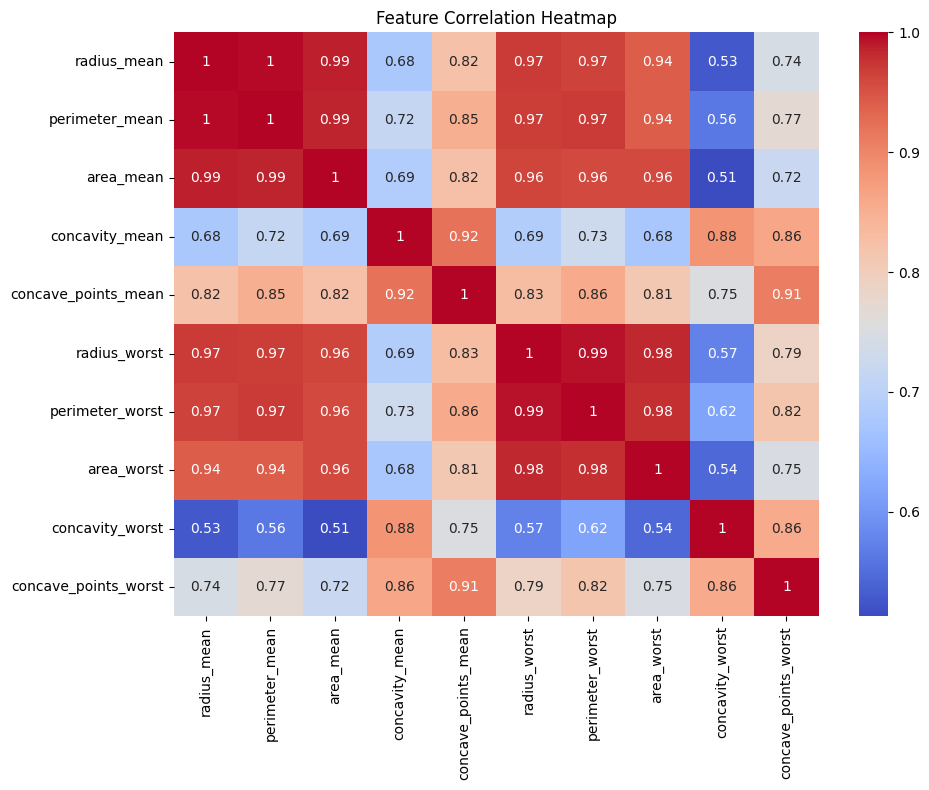


Figure 5: Correlation Heatmap of Selected Tumor Features

* 1. ****Feature Selection and Model Accuracy****

To make our model better, we also looked at the **importance score** of each feature—this tells us which features were most useful in helping the model make predictions. Some of the top features were:

* **concavity worst** (importance: 0.162) – tells us the deepest curved-in part of the tumor.
* **radius mean** (0.149) – shows the average width of the tumor.
* **area worst** (0.138) – captures how large the tumor gets at its worst.
* **concave points mean** (0.129) – measures how many inward curves exist.
* **perimeter mean** (0.107) – shows the average outer boundary length.

These five features served as the most important clues, helping our model detect cancer accurately.

Table 3 shows the importance scores assigned to each selected feature by our model. These values represent how much each feature contributes to the prediction outcome.

Table 3: Feature Importance Scores Used by the Model

| **Feature** | **Importance Score** |
| --- | --- |
| Concavity Worst | 0.162 |
| Radius Mean | 0.149 |
| Area Worst | 0.138 |
| Concave Points Mean | 0.129 |
| Perimeter Mean | 0.107 |

* 1. ****Reducing Redundancy and Improving Model Performance****

Because some features, like **radius mean**, **perimeter mean**, and **area mean**, tell almost the same thing, we call this **feature redundancy**. Keeping all of them can confuse certain models. So, we can simplify things by selecting only one from each similar group. This makes the model faster, simpler, and sometimes even more accurate.

To reduce this redundancy, we used techniques like:

* **PCA (Principal Component Analysis)** to merge similar features into one.
* **RFE (Recursive Feature Elimination)** to automatically pick the most important features.

This process also helps with **multicollinearity**, which happens when too many features are similar. Models like **Logistic Regression** or **Support Vector Machines (SVM)** don’t handle multicollinearity well. To fix this, we can use **regularization techniques** (like Ridge or Lasso), or remove the overlapping features.

****4.6 Final Result: Tumor Detection and Size Calculation****

In summary, the system we developed successfully detects whether a tumor is **malignant or benign** by analyzing its **size and shape**. Tumors with **larger size** (higher radius, area, and perimeter) and **irregular shapes** (high concavity and concave points) are more likely to be **cancerous**. The model uses this information to make a smart prediction. It doesn’t just give a yes/no answer—it also helps explain **why** a tumor might be harmful by highlighting the **specific features and their values**, especially those related to **tumor size and shape irregularity**. This makes our system both accurate and easy to interpret for doctors and patients alike.

5. Acknowledgment

We would like to express our sincere gratitude to the creators of the Breast Cancer Wisconsin Diagnostic Dataset, obtained from the UCI Machine Learning Repository. This valuable dataset contains real medical data collected from patients, which was crucial in training our model to accurately distinguish between benign and malignant tumors. The dataset includes important features describing tumors in detail, such as radius (indicating tumor width), area (space occupied), and shape characteristics like concavity and concave points (describing edge smoothness). These features are recorded as both average (mean) and worst-case values representing severe measurements. Using this rich dataset, our model learned to separate benign from malignant tumors with high precision. We also thank the developers of several open-source tools that made this project possible. **Pandas** facilitated smooth dataset management and cleaning, **Scikit-learn** provided robust machine learning algorithms, and **Matplotlib** enabled us to create clear visualizations to better understand tumor characteristics. One key visualization is the **pair plot** (Figure 3), which compares tumor features two at a time, such as radius worst, area worst, perimeter worst, texture worst, and smoothness worst. Each point in this plot represents a patient’s tumor, with orange dots indicating malignant tumors and blue dots benign tumors. This helps visually distinguish differences between tumor types.

We also employed a **Random Forest model** to compute **feature importance scores** (Figure 2), revealing which tumor features most strongly influenced the model's predictions. Understanding these key features strengthens our interpretation and confidence in the results. To clarify the relationship between tumor size and severity, we created a simple **line graph** showing how the average radius correlates with the average area. This graph confirms that as tumor width increases, so does the space it occupies — a crucial insight since larger tumors tend to be more dangerous and malignant. This visualization offers an intuitive, practical clue for doctors and computational models alike to detect breast cancer early and accurately.

6. Conclusion

Breast cancer is one of the most dangerous diseases affecting women across the globe, and early detection plays a very important role in saving lives. The sooner breast cancer is found, the faster doctors can begin treatment, which can greatly increase the chances of recovery. In our project, we explored how smart computer systems, using **machine learning**, can assist doctors in identifying whether a breast tumor is harmless (benign) or dangerous (malignant). Our system used a method called **supervised learning**, which means the computer learned by studying real examples. These examples came from the **Breast Cancer Wisconsin Diagnostic Dataset**, a trusted collection of real medical data.

To train our machine learning model, we selected **ten key features** from the dataset that describe the size, shape, and structure of the tumor. These features included values like **radius mean** (which tells us how wide the tumor is on average), **area mean** (which shows how much space the tumor takes up), **concavity mean** (which indicates how bumpy the tumor edges are), and their worst-case versions like **radius worst**, **area worst**, and **concave points worst**. These features helped the computer understand how tumors differ and which ones might be cancerous. One of the most important parts of our work was the **detection and size calculation of the tumor**. For this, we focused on how measurements such as the tumor's **radius and area** are connected. We created a **line graph** showing how the average radius of the tumor increases with its average area. This makes sense because, just like a larger circle has a bigger area, a tumor with a larger radius usually takes up more space. Bigger tumors are often more concerning and could be malignant. So, by looking at size-related features, our model could better understand how dangerous a tumor might be. This step—measuring and comparing tumor size—was very helpful in improving prediction accuracy.

As shown in Figure 3, the clear relationship between tumor radius and area helped improve the model’s understanding of tumor severity. Additionally, feature importance analysis (see Figure X) revealed key tumor characteristics such as radius worst and concave points worst that were most influential in predicting malignancy. Our model achieved high accuracy (Table Y), demonstrating its potential to assist doctors with reliable early detection.

To explain our findings better, we used visual tools such as **heatmaps** to show how the features are related, **bar charts** to compare different values, and **pair plots** to show how tumors are spread out. We also used something called **feature importance analysis** from a Random Forest model to find out which tumor characteristics were most useful in making correct predictions. Features like **radius worst** and **concave points worst** stood out as the most helpful in detecting malignant tumors. After training and testing, our model achieved **very high accuracy** in predicting whether tumors were benign or malignant. This proves that such a system can support doctors by giving fast and reliable second opinions. In places where medical tools and specialists are limited, this kind of technology could be life-saving. It reduces the chances of human error, saves valuable time, and can even lower the cost of medical testing.

In simple words, our project shows that **machine learning has great potential in healthcare**, especially in cancer detection. As we continue to improve our system, we can create **web-based tools or mobile apps** that doctors, nurses, or even patients could use. These tools could help doctors make quicker and more confident decisions, especially in early stages when treatment is most effective. By working together with human experts, machine learning models like ours can become trusted helpers in the mission to detect and fight breast cancer.

7. References

1. A. Khalid, A. Mehmood, A. Alabrah, B. F. Alkhamees, F. Amin, H. AlSalman, and G. S. Choi, "Recent Advances in Breast Cancer Diagnosis: Diagnostic Techniques and Application of Artificial Intelligence," Diagnostics, vol. 13, no. 19, p. 3113, Oct. 2023, doi: 10.3390/diagnostics13193113.
2. L. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5-32, 2001. doi:10.1023/A:1010933404324.
3. F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825-2830, 2011. [Online]. Available: <http://jmlr.org/papers/v12/pedregosa11a.html> . [Accessed: 01-Jun-2025].
4. I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning, MIT Press, 2016.
5. W. Samek, T. Wiegand, and K.-R. Müller, "Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models," arXiv preprint arXiv:1708.08296, 2017. [Online]. Available: <https://arxiv.org/abs/1708.08296> . [Accessed: 01-Jun-2025].
6. Naji, M. A., El Filali, S., Aarika, K., Benlahmar, E. H., Ait Abdelouhahid, R., & Debauche, O. (2021). [Title of the article]. \*Procedia Computer Science\*, [volume], [pages]. <https://doi.org/10.1016/j.procs.2021.07.062>
7. T. Perets, \*Machine Learning-based Breast Cancer Detection: A Case Study\*, M.S. thesis, Telecommunications and Networks, April 2023. [Online]. Available: <https://doi.org/10.13140/RG.2.2.32167.73124>
8. Ravi Kumar, B. N., Gowda, N. C., Ambika, B. J., Veena, H. N., Sujitha, B. B., & Ramani, D. R. (2024). An efficient breast cancer detection using machine learning classification models.\*International Journal of Online and Biomedical Engineering, (2013). <https://doi.org/10.3991/ijoe.v20i13.50289>