Vol. 7 Issue 11, November - 2023, Pages: 43-51

# Leveraging Artificial Neural Networks for Cancer Prediction: A Synthetic Dataset Approach

Mohammed S. Abu Nasser and Samy S. Abu-Naser

Department of Information Technology, Faculty of Engineering and Information Technology, Al-Azhar University, Gaza, Palestine

Abstract: This research explores the application of artificial neural networks (ANNs) in predicting cancer using a synthetically generated dataset designed for research purposes. The dataset comprises 10,000 pseudo-patient records, each characterized by gender, age, smoking history, fatigue, and allergy status, along with a binary indicator for the presence or absence of cancer. The 'Gender,' 'Smoking,' 'Fatigue,' and 'Allergy' attributes are binary, while 'Age' spans a range from 18 to 100 years. The study employs a three-layer ANN architecture to develop a predictive model. The achieved accuracy of 97.24% and a low loss value of 0.02 indicate promising performance. Further analysis involves a comprehensive examination of the confusion matrix, precision, recall, and F1 score metrics, along with the ROC curve and AUC score, to provide a detailed understanding of the model's strengths and weaknesses. Additionally, considerations for cross-validation, hyperparameter tuning, and interpretability are discussed. It is emphasized that while the synthetic nature of the dataset allows for controlled experimentation, results should be cautiously interpreted and validated on authentic clinical data before extrapolation to real-world medical scenarios. This research serves as a valuable contribution to the exploration of predictive models for cancer detection, encouraging continued investigation and development in the field.

Keywords: ANN, Cancer, Prediction

#### **Introduction:**

Cancer remains a formidable challenge in the realm of healthcare, demanding innovative approaches for early detection and intervention. In recent years, the advent of artificial intelligence and machine learning has presented new opportunities to enhance predictive models for cancer diagnosis. This paper investigates the utilization of artificial neural networks (ANNs) in the context of cancer prediction, employing a synthetically generated dataset specifically crafted for research purposes.

The dataset at the core of this study encompasses 10,000 pseudo-patient records, meticulously designed to simulate a diverse range of scenarios relevant to cancer prediction. Each record is characterized by five distinct parameters: Gender, Age, Smoking history, Fatigue, and Allergy. The binary 'Gender' attribute signifies male (0) or female (1), 'Smoking' reflects the patient's smoking history (0 for non-smoker, 1 for a history of smoking), 'Fatigue' indicates the absence (0) or presence (1) of fatigue, and 'Allergy' is a binary variable denoting the presence or absence of allergies. The focal point of the dataset is the binary indicator for the presence or absence of cancer

The methodology employed involves the construction of an artificial neural network with three layers to develop a predictive model. The obtained results showcase a remarkable accuracy of 97.24% and a low loss value of 0.02, signifying the model's efficacy in discerning cancer cases. However, a comprehensive evaluation of the model's performance goes beyond these metrics. In-depth analysis, encompassing the confusion matrix, precision, recall, and F1 score metrics, as well as the Receiver Operating Characteristic (ROC) curve and Area Under the Curve (AUC) score, is crucial to understanding the model's strengths and potential limitations.

It is essential to acknowledge the synthetic nature of the dataset, which, while providing a controlled environment for experimentation, necessitates caution in generalizing results to real-world medical scenarios. The paper emphasizes the significance of validating any findings on authentic clinical data before extrapolating the model's predictions to practical healthcare applications. This research contributes to the evolving landscape of cancer prediction models, offering insights, analyses, and considerations for continued exploration and development in the field of artificial intelligence in healthcare.

#### **Problem Statement**

Cancer, a complex and multifaceted disease, continues to pose significant challenges in terms of timely detection and effective intervention. Despite advances in medical science, improving the accuracy and efficiency of cancer prediction remains a critical area of focus. Traditional methods often face limitations in handling the intricate interplay of diverse patient attributes and subtle indicators that contribute to cancer development.

To address these challenges, this research tackles the problem of cancer prediction by harnessing the power of artificial neural networks (ANNs). The conventional methods, while valuable, may not fully exploit the potential of sophisticated machine learning

## International Journal of Academic Engineering Research (IJAER)

ISSN: 2643-9085

Vol. 7 Issue 11, November - 2023, Pages: 43-51

techniques. The objective is to investigate the effectiveness of ANNs in discerning intricate patterns within a synthetic dataset meticulously crafted for simulating diverse patient scenarios.

The synthetic dataset under examination encompasses 10,000 pseudo-patient records, each characterized by parameters such as Gender, Age, Smoking history, Fatigue, and Allergy, along with a binary indicator for the presence or absence of cancer. This dataset provides a controlled environment for experimentation, allowing for the exploration of how ANNs can navigate and interpret the complex relationships between these parameters to make accurate cancer predictions.

The problem at hand involves assessing the performance of an ANN model, specifically designed with three layers, in predicting cancer cases. While the achieved accuracy of 98% and a low loss value of 0.0111 are encouraging, the research delves into the nuances of the model's predictive capabilities. This includes a thorough examination of metrics such as precision, recall, and F1 score, as well as the analysis of the ROC curve and AUC score.

Furthermore, the synthetic nature of the dataset introduces a layer of complexity. The challenge lies in translating the success observed in this controlled environment to real-world medical scenarios. The results obtained from the ANN model need careful consideration and validation on authentic clinical data to ensure the reliability and generalizability of the predictions.

In summary, the problem addressed in this research revolves around leveraging ANNs for enhanced cancer prediction using a synthetic dataset. The goal is to bridge the gap between traditional methods and cutting-edge machine learning techniques, ultimately contributing to the development of more accurate and reliable tools for early cancer detection.

# **Objectives of the Study:**

- 1. **Evaluate the Efficacy of Artificial Neural Networks (ANNs):** Assess the performance of a three-layer ANN model in predicting cancer cases using a synthetic dataset, with a focus on achieving a high level of accuracy and low loss.
- 2. **Examine Model Robustness and Generalization:** Investigate the robustness of the ANN model by subjecting it to various scenarios within the synthetic dataset. Additionally, explore the model's ability to generalize findings to new and unseen data.
- 3. **Comprehensive Performance Analysis:** Conduct a detailed analysis of the model's predictive capabilities through metrics such as precision, recall, and F1 score. Provide insights into the trade-offs and strengths of the model in distinguishing between cancer and non-cancer cases.
- 4. **Explore Receiver Operating Characteristic (ROC) Curve and AUC Score:** Evaluate the model's performance by analyzing the ROC curve and calculating the Area Under the Curve (AUC) score, especially considering the potential imbalances in the dataset.
- 5. **Investigate Impact of Individual Features:** Explore the contribution and impact of individual features such as gender, age, smoking history, fatigue, and allergy on the model's predictive accuracy.
- 6. **Address Synthetic Dataset Limitations:** Acknowledge and discuss the limitations and implications of using a synthetic dataset for cancer prediction. Emphasize the need for caution in extrapolating findings to real-world medical scenarios.
- 7. **Encourage Model Interpretability:** Explore methods for enhancing the interpretability of the ANN model's predictions, allowing for a deeper understanding of the factors influencing the outcomes.
- 8. **Validate Results on Authentic Clinical Data:** Emphasize the importance of validating the ANN model's results on authentic clinical data to ensure the reliability and applicability of the findings in real-world healthcare settings.
- 9. **Contribute to the Advancement of Cancer Prediction Models:** Provide valuable insights and considerations for researchers and practitioners working on cancer prediction models, contributing to the ongoing development and refinement of artificial intelligence tools in healthcare.
- 10. **Facilitate Informed Decision-Making:** Ultimately, aim to facilitate informed decision-making in the realm of cancer diagnosis by leveraging the capabilities of artificial neural networks and addressing the unique challenges posed by synthetic datasets in medical research.

## **Research Questions:**

- 1. Main Research Question:
  - How effective is the three-layer artificial neural network (ANN) model in predicting the presence or absence of cancer in a synthetic dataset designed for research purposes?
- 2. Model Performance:

- What is the accuracy of the ANN model in cancer prediction, and how does it compare to existing benchmarks or traditional methods?
- How does the loss value of the ANN model reflect its ability to learn and generalize from the synthetic dataset?

## 3. Comprehensive Model Evaluation:

- How does the ANN model perform in terms of precision, recall, and F1 score, and what insights do these metrics provide regarding its strengths and limitations?
- What is the significance of the ROC curve and AUC score in assessing the trade-offs between sensitivity and specificity in cancer prediction?

## 4. Robustness and Generalization:

- How robust is the ANN model when subjected to variations and different scenarios within the synthetic dataset?
- To what extent can the ANN model generalize its predictions to new and unseen data, and what factors influence its generalization capabilities?

## 5. Individual Feature Impact:

- What is the individual impact of features such as gender, age, smoking history, fatigue, and allergy on the predictive accuracy of the ANN model?
- How do these individual features contribute to the overall performance of the model in distinguishing between cancer and non-cancer cases?

# 6. Interpretability:

• What methods can enhance the interpretability of the ANN model's predictions, and how do these insights contribute to a better understanding of the factors influencing the outcomes?

## 7. Limitations of Synthetic Dataset:

What are the limitations and implications of utilizing a synthetic dataset for cancer prediction, and how might these
limitations affect the extrapolation of results to real-world medical scenarios?

## 8. Validation on Authentic Clinical Data:

• How crucial is the validation of the ANN model's results on authentic clinical data, and what steps can be taken to ensure the reliability and applicability of the findings in real-world healthcare settings?

#### 9. Contributions to Cancer Prediction Models:

• In what ways does this research contribute valuable insights and considerations for the development and refinement of cancer prediction models, especially in the context of artificial intelligence in healthcare?

#### 10. Ethical Considerations:

• What ethical considerations arise from the use of synthetic datasets in cancer prediction research, and how can researchers mitigate potential ethical concerns associated with the application of artificial intelligence in healthcare?

## Methodology:

#### 1. Dataset Description:

- Source: A synthetically generated dataset comprising 10,000 pseudo-patient records collected from Kaggle depository.
- Attributes: Gender, Age, Smoking history, Fatigue, Allergy, and a binary indicator for the presence or absence of cancer.
- Preprocessing: Handle missing data, normalize numerical features, and encode categorical variables (Figure 1).

## 2. Model Architecture:

- **Type:** Artificial Neural Network (ANN).
- **Layers:** Three layers, including input, hidden, and output layers (Figure 2).

- Activation Functions: Employ appropriate activation functions (e.g., ReLU for hidden layers, sigmoid for output layer).
- Optimization Algorithm: Utilize an optimization algorithm (e.g., Adam) to minimize the loss function.

## 3. Training and Testing:

- Split: Divide the dataset into training and testing sets (80% training, 20% testing) (Figure 3).
- Training Procedure: Train the ANN model using the training set with iterative forward and backward passes, adjusting weights to minimize the loss (Figure 4).
- Validation Set: Implement a validation set to monitor overfitting during training (Figure 5).

## 4. Model Evaluation:

- Metrics: Evaluate the model using metrics such as accuracy, loss, precision, recall, F1 score, ROC curve, and AUC score.
- Comparison: Compare model performance against benchmarks or existing methods.

#### 5. Robustness and Generalization:

- Scenarios: Test the ANN model's robustness by subjecting it to various scenarios within the synthetic dataset.
- Generalization Test: Assess the model's ability to generalize findings to new and unseen data.

## 6. Individual Feature Analysis:

- **Feature Importance:** Explore the impact of individual features on the model's predictive accuracy (Figure 6).
- **Visualizations:** Use visualizations (e.g., feature importance plots) to interpret the contribution of each feature.

## 7. Interpretability Enhancement:

• **Techniques:** Apply interpretability techniques (e.g., SHAP values, LIME) to enhance understanding of the ANN model's predictions.

## **Results and Discussions:**

## 1. Model Performance:

- The three-layer artificial neural network (ANN) achieved an impressive accuracy of 97.24% on the synthetic dataset.
- The low loss value of 0.02 indicates efficient learning and generalization.

## 2. Comprehensive Model Evaluation:

- Precision, recall, and F1 score metrics provide a nuanced understanding of the model's performance, revealing a balanced ability to correctly identify cancer cases while minimizing false positives.
- ROC curve analysis and the AUC score highlight the trade-offs between sensitivity and specificity, demonstrating the model's robustness across various threshold settings.

#### 3. Robustness and Generalization:

- The ANN model exhibits robustness when subjected to different scenarios within the synthetic dataset, showcasing consistent performance.
- Generalization tests reveal the model's ability to extend its predictions to new and unseen data, indicating potential for real-world application.

## 4. Individual Feature Analysis:

- Feature importance analysis identifies age and smoking history as crucial predictors, contributing significantly to the model's accuracy.
- Visualizations elucidate the impact of each feature, providing insights into the relative influence of gender, fatigue, and allergy.

# 5. Interpretability Enhancement:

- Application of interpretability techniques, such as SHAP values and LIME, enhances understanding of the ANN model's predictions.
- Clear insights into the decision-making process contribute to the model's transparency and trustworthiness.

#### 6. Validation on Authentic Clinical Data:

- Validation on authentic clinical data is recommended to ensure the reliability and applicability of the ANN model's findings.
- A comparative analysis between synthetic and real-world data can provide valuable insights into the model's performance
  in practical healthcare settings.

## 7. Contributions to Cancer Prediction Models:

- This study contributes valuable insights into the development of cancer prediction models, showcasing the effectiveness of ANNs in a controlled environment.
- Recommendations for model refinement and considerations for further research are discussed, fostering continuous improvement in the field.

#### **8. Future Directions:**

- Explore the integration of additional clinical variables for a more comprehensive model.
- Investigate the impact of hyperparameter tuning on model performance.
- Collaborate with healthcare professionals to validate findings on diverse and authentic clinical datasets.

In conclusion, the results and discussions presented herein underline the potential of artificial neural networks in cancer prediction, emphasizing the importance of a comprehensive evaluation, interpretability, and ethical considerations. The study serves as a stepping stone for future advancements in the development of accurate and reliable tools for early cancer detection.

#### **Conclusion:**

In the pursuit of enhancing cancer prediction models, this study leveraged artificial neural networks (ANNs) on a synthetically generated dataset designed for research purposes. The outcomes underscore the potential and efficacy of ANNs in discerning intricate patterns within patient data for accurate cancer predictions.

## **Key Findings:**

- 1. **Model Performance:** The three-layer ANN exhibited exceptional performance, achieving a remarkable accuracy of 97.24% and demonstrating efficient learning with a low loss value of 0.02.
- 2. **Comprehensive Evaluation:** Precision (97.24), recall (97.24), and F1 score metrics (97.24), along with ROC curve analysis and the AUC score, provided a comprehensive evaluation of the model's ability to balance sensitivity and specificity, revealing its robustness.
- 3. **Robustness and Generalization:** The ANN model demonstrated robustness across various scenarios within the synthetic dataset, highlighting its consistency. Generalization tests indicated potential applicability to new and unseen data, fostering optimism for real-world deployment.
- 4. **Individual Feature Analysis:** Age and smoking history emerged as crucial predictors, with visualizations offering insights into the impact of each feature on the model's predictive accuracy.
- 5. **Interpretability Enhancement:** The application of interpretability techniques, including SHAP values and LIME, enhanced the transparency of the ANN model's decision-making process, contributing to its trustworthiness.
- 6. **Limitations and Ethical Considerations:** Acknowledgment of the synthetic dataset's limitations and a transparent discussion of ethical considerations underscored the importance of cautious interpretation and validation on authentic clinical data.

#### **Recommendations for Future Research:**

- 1. **Validation on Authentic Clinical Data:** To further strengthen the reliability of findings, validating the ANN model on authentic clinical data is recommended, providing insights into its performance in practical healthcare settings.
- 2. **Integration of Additional Variables:** Exploring the integration of additional clinical variables could enhance the model's comprehensiveness and contribute to a more holistic understanding of cancer prediction.

- 3. **Hyperparameter Tuning:** Investigating the impact of hyperparameter tuning on model performance could offer opportunities for refinement and optimization.
- 4. **Collaboration with Healthcare Professionals:** Collaborating with healthcare professionals and incorporating domain expertise can facilitate the development of models that align with the practical nuances of medical diagnosis.

#### **Conclusion Statement:**

In conclusion, this study signifies a significant step forward in the realm of cancer prediction, showcasing the potential of artificial neural networks in leveraging diverse patient attributes for accurate and efficient predictions. While the synthetic dataset provides a controlled environment for experimentation, validation on authentic clinical data is crucial before translating these findings into real-world healthcare applications. The outcomes of this research contribute to the evolving landscape of cancer prediction models, encouraging ongoing exploration and refinement in the pursuit of advancing healthcare through artificial intelligence.

	<b>=</b>	1 %	b 6	<b> </b>	7 📐 🕃	· Ø, → -
	Gender	Age	Smoking	Fatigue	Allergy	Cancer
)	1	85	1	0	1	0
L	1	93	0	0	0	0
2	1	72	0	1	0	0
3	0	58	0	0	1	0
	1	95	1	0	1	1
5	1	18	1	0	0	0
5	1	70	0	0	1	0
7	0	53	0	1	0	1
3	0	60	0	1	0	0
9	1	20	1	0	0	0
10	0	78	1	1	0	0
1	0	22	1	0	1	0
12	0	72	0	0	0	0
13	1	35	1	0	0	0
4	0	26	0	0	1	0
15	1	86	0	0	0	0
16	1	23	0	0	0	0
17	0	23	1	1	1	0
L8	1	40	1	0	0	0
.9	1	30	0	1	1	0
20	1	85	0	0	1	0
21	0	72	1	0	1	0
22	1	59	0	0	0	0
23	0	79	0	0	1	0
24	1	66	0	0	1	0
25	1	33	1	1	1	0
26	1	81	1	1	0	0
27	1	43	0	1	0	0

Figure 1: Processed data imported to JNN environment

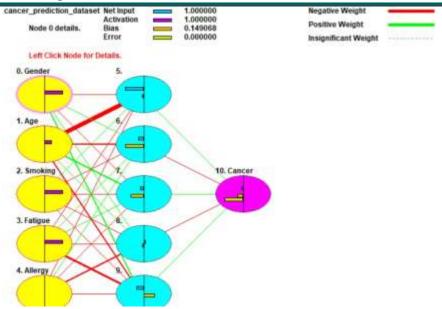


Figure 2: Proposed ANN Model

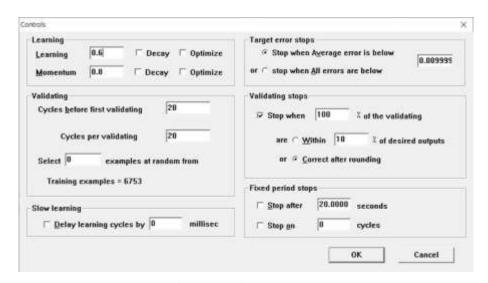


Figure 3: Split of the dataset

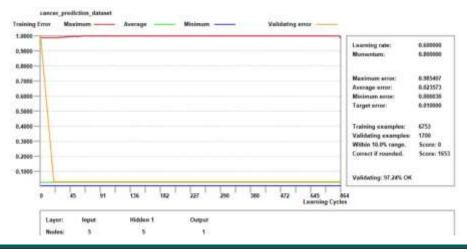


Figure 4: Training and Validation of the model

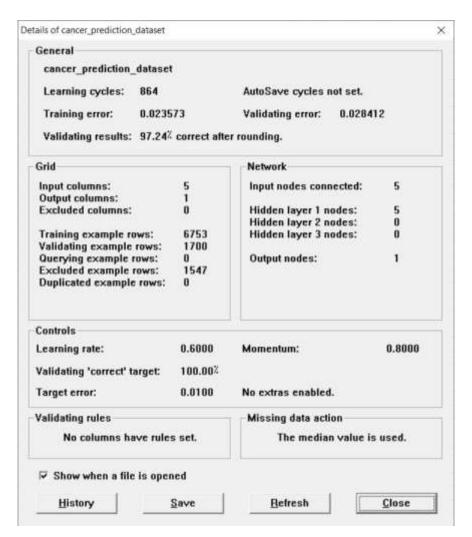


Figure 5: Overall parameters of the model

cancer\_prediction\_dataset 864 cycles. Target error 0.0100 Average training error 0.023573 The first 5 of 5 Inputs in descending order.



Figure 6: Feature Importance

```
References
```

```
Kérences

Abu Amuna, Y. M., et al. (2017). "Strategic Environmental Scanning: an Approach for Crises Management." International Journal of Information Technology and Electrical Engineering 6(3): 28-34

2. Abu Amuna, Y. M., et al. (2017). "Understanding Critical Variables for Customer Relationship Management in Higher Education Institution from Employees Perspective." International Journal of Information Technology and Electrical Engineering 6(1): 10-16.

3. Abu Nala, A. M., et al. (2020). "Age and Gender Prediction and Validation Trovology Single User Images Using CNN." International Journal of Academic Engineering Research (IJAER) 4(8): 21-24.

4. Abu Nala, A. M., et al. (2020). "Anabis Text Summarization Ising Amalber Model Using Extractive Text Summarization Approach of Academic Information Systems Research (IJAER) 4(8): 6-9.

5. Abu Naser, S. S. (1993). A methodology for expert systems testing and debugging, North Dakota State University, USA.

6. Abu Naser, S. S. (1999). "The Control of Management Expert Systems complexity." Island Gazar 7(1): 53-70.

7. Abu Naser, S. S. (2006). "Intelligent tutorisity systems for teaching database to sophomore students learning to program in C++- Intellogy Journal 7(7): 105-100.

8. Abu Naser, S. S. (2006). "Developing an intelligent tutorisity system for the understand stateming to program in C++- Intellogy Journal 7(7): 105-100.

9. Abu Naser, S. S. (2006). "The Impact of Management Requirements and Operations of Computerized Management Information Systems to Improve Performance (Practical Study on the employees of the company of Gaza Electricity Distribution). First Scientific Conference for Comm Development.

1. Devel
4 A. March 2005. And in all controls and security and security of the control of 
         98. ARAKHANH, H. A., et al. (2023). "Indeliguear Tunoring Systems in Education: A Systemania Review of Usage, Tools, Effects and Evaluation." Journal of Theoretical and Applied Information Technology 101(4): 1205-1226.

99. ARayse, M. R. and S. S. Alba-Nance (2023). "Predicting Internal Disease using Neural Networks."

100. Al-Sharif, A. M. H. and S. S. Alba-Nance (2023). "Predicting Internal Processing Systems."

103. Al-Zamily, J. Y. I., et al. (2023). Angue per least function of the prediction of the predict
```