Journal of Theoretical and Applied Information Technology
15th August 2025. Vol.103. No.15

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

A UNIFIED CNN-BASED FRAMEWORK FOR GENERALIZED AND REAL-TIME COLORECTAL CANCER PREDICTION AND DIAGNOSIS: BRIDGING DATA GAPS, ENHANCING INTERPRETABILITY, AND PERSONALIZING OUTCOMES

K YOGESWARA RAO[1], S ADI NARAYANA[2]

[1] Associate Professor, Department of Computer Science and Engineering, GITAM School of Technology, GITAM (Deemed to be University), Visakhapatnam, India
[2] Professor, Department of CSE (Data Science), Anil Neeru Konda Institute of Technology and Sciences Sangivalasa. Visakhapatnam, India
E-mail: [1]yogiindusisu@gmail.com, [2] sadinarayana.csd@anits.edu.in

ABSTRACT

Colorectal cancer (CRC) remains one of the most common and deadly cancers worldwide, making early detection and accurate diagnosis more important than ever. In this work, we introduce UniCRC-Net—a smart, CNN-based system designed to predict and diagnose colorectal cancer in real time, using structured patient data. Unlike many current machine learning and deep learning models, which struggle with scattered data, lack of explainability, and generic predictions, this unified approach brings together multiple patient details—like age, gender, pathology scores, gene markers, diet, and environment—into a streamlined and intelligent framework. The model is trained on a carefully constructed synthetic dataset and optimized using the Adam algorithm over 50 training epochs. It performs exceptionally well, hitting a perfect 100% accuracy, F1-score, and AUC, which means it's both highly precise and consistent in identifying cancer cases. The results are backed by clear visualizations—such as accuracy and loss graphs, a confusion matrix, and a sharp ROC curve—demonstrating how stable and dependable the model is throughout its training. What sets UniCRC-Net apart is its real-time capability, its ability to personalize predictions, and its transparent design, which makes it easier to trust in clinical use. It's also built with the future in mind—ready for integration with federated learning systems that protect patient privacy while enabling collaboration across hospitals and regions. In short, this framework not only fills major gaps in CRC diagnostics but also moves us a step closer to AI-powered, patient- specific cancer care that's fast, secure, and clinically meaningful.

Keywords: *Colorectal Cancer, Deep Learning, Convolutional Neural Network, Real-Time Diagnosis, Personalized Healthcare, ROC-AUC, Clinical Decision Support and Interpretability in AI*

1. INTRODUCTION

1.1 Background on Colorectal Cancer (CRC) Prevalence and Impact

Colorectal cancer (CRC) is a serious global health challenge. It's consistently among the top three most commonly diagnosed cancers and remains a leading cause of cancer-related deaths around the world. According to the World Health Organization, millions of people are diagnosed with CRC every year, and it affects both wealthy and lower-income countries alike. Several factors contribute to its widespread nature—aging populations, unhealthy diets, sedentary lifestyles, and even family history all play a role. What makes CRC especially concerning is that it tends to develop slowly over time, meaning that with the right tools, it can often be caught early and treated effectively. But the impact goes beyond health alone. CRC brings with it economic strain, higher medical expenses, loss of productivity, and significant emotional stress for both patients and their families. Despite ongoing awareness efforts, many cases are stil detected too late, largely because early symptoms are mild—or even nonexistent—and access to timely screening is still limited in many parts of the world. Despite awareness efforts and the availability of screening tools, many cases are detected at an advanced stage—primarily due to symptom-free early phases and limited access to diagnostics in underserved regions. This highlights the urgent need for faster, more affordable, and easily accessible screening and diagnostic tools that can reach people everywhere—regardless of location or resources.

Figure-1 gives a clear picture of how colon cancer develops step by step, starting from normal, healthy tissue and gradually turning into a serious, life-threatening condition. It begins at Stage 0, also called carcinoma in situ, where abnormal cells are found only on the inner lining of the colon. At this point, it's highly treatable because the cancer hasn't yet spread deeper. As it progresses to Stages I and II, the cancer starts to move into deeper layers of the colon wall, often causing symptoms like changes in bowel habits or light bleeding. In Stage III, the cancer spreads to nearby lymph nodes, which means it's becoming more aggressive and usually needs surgery plus chemotherapy. Finally, in Stage IV, it reaches distant organs like the liver or lungs, making treatment much more difficult and reducing survival chances significantly. The figure not only outlines how colon cancer evolves but also helps doctors decide on treatment plans. Even more, it provides a solid foundation for building AI-powered tools that can predict how the disease will progress—using a mix of personalized data like patient history, biomarkers, and imaging results.

1.2 Limitations of Traditional Diagnosis Methods

When it comes to detecting colorectal cancer (CRC), doctors usually rely on tests like colonoscopy, sigmoidoscopy, fecal occult blood tests (FOBT), and CT scans. While colonoscopy is considered the gold standard, it's also invasive, costly, and demands a lot of resources—making it harder to access, especially in places with limited medical infrastructure. On top of that, the discomfort and preparation involved often lead people to avoid regular screening, even when it's recommended. Studies have pointed out several downsides to these traditional methods, such as missed polyps, differences in how doctors interpret results, and the complexity of analyzing tissue samples. As noted by Tharwat et al. [1], although these tests can be effective, they ren't well-suited for fast, scalable, or automated diagnosis. That's why there's a growing need for smarter, AI-driven tools that can make screening quicker, easier, and more accessible to everyone.

1.3 Need for AI-Driven, Personalized, Real-Time Diagnostic Systems

Artificial Intelligence (AI)—especially deep learning tools like Convolutional Neural Networks (CNNs)—is changing how we approach medical diagnosis today. In the case of colorectal cancer (CRC), AI can analyze a wide range of clinical data, from pathology images to genetic profiles and even lifestyle habits, helping doctors make faster and more accurate decisions. What's amazing is that CNNs have shown diagnostic accuracy equal to, or even better than, experienced pathologists [2][3]. Today's cancer care is all about personalization. Things like a person's diet, inflammation levels, genetic risks, and environmental factors all influence how CRC develops and progresses. AI systems that take these into account can provide customized diagnosis and treatment suggestions. Research by Kalpana et al. [3] and Mansur et al. [4] shows that AI not only improves prediction but also makes results more understandable for clinicians—and can save valuable time in urgent clinical situations.

1.4 Chapter Objective and Structure

In this chapter, we introduce UniCRC-Net, a smart and streamlined CNN-based system designed to predict and diagnose colorectal cancer (CRC) in real-time using structured clinical data. This framework was built to solve some big problems found in current diagnostic tools— like fragmented data, lack of personalization, hard-to-interpret results, and slow response times. By training a compact neural network on a carefully designed synthetic dataset, the model achieved outstanding results, including perfect accuracy, AUC, F1-score, and super-fast training. Here's how the chapter is structured: Section 2 reviews recent machine learning (ML) and deep learning (DL) studies on CRC detection, covering both strengths and limitations. Section 3 dives into how the model works—its data pipeline, network layers, and optimization process. Section 4 presents the experimental outcomes with clear visualizations and performance metrics. In Section 5, we interpret the results, compare them with past research, and look at how this system could be used in real clinics. Finally, Section 6 wraps up with key takeaways and future directions to bring this framework closer to real-world medical deployment.

2. LITERATURE REVIEW

2.1 Evolution of ML/DL Approaches in Colorectal Cancer Detection

In the last ten years, machine learning (ML) and deep learning (DL) have made big waves in how we detect and diagnose colorectal cancer (CRC). These smart technologies help doctors automatically analyze complex medical data—like tissue scans, genetic profiles, and patient health records—much faster and often more accurately than traditional methods. Models like Convolutional Neural Networks (CNNs) and Support Vector Machines (SVMs) have proven to be incredibly effective at spotting patterns and predicting outcomes across different types of data. Research using open datasets like TCGA and COAD has shown that deep learning tools can even outperform some of the standard diagnostic processes. Still, there are a few roadblocks. Many AI models struggle with things like imbalanced datasets, overfitting, and being hard to interpret. Plus, it's not always easy to use them in real-time clinical settings. So while the progress is exciting, more work is needed to make these tools widely usable in hospitals and clinics.

2.2 Contributions and Limitations in Existing Research

Even though machine learning and deep learning have shown great promise in detecting colorectal cancer (CRC), there are still some major challenges. Many of the current models don't scale well or generalize across different populations, and most are tested only on data from a single hospital—which limits how useful they are elsewhere. Another big issue is that many systems don't combine different types of patient data (like medical history, lifestyle, and genetics), even though this kind of fusion is key for accurate, personalized medicine. On top of that, a lot of deep learning models act like "black boxes"—they give you a result, but it's hard to explain how they got there. That makes it tough for doctors to trust them, and it's also a barrier when it comes to getting regulatory approval. Important patient information like long- term health history and daily habits is often left out, too. To move forward, we need transparent, explainable AI models that are trained on diverse and realistic datasets—models that doctors can trust and that work for everyone, not just specific groups.

2.3 Analysis by Tharwat et al. (2022)

Tharwat and colleagues [1] did a deep dive into how machine learning and deep learning models are being used to analyze colon cancer through histopathology images. They explored various algorithms—like CNNs, Random Forests, and SVMs—most of which were trained using well-known datasets like TCGA. While the review gave great insight into how these models work with different types of data, it also pointed out a major gap: most of the models weren't built for real-time use or flexible enough to handle the variety of data found in real clinical settings. The review emphasized the need for smarter, more adaptable AI systems that can work with different types of patient data and help automate the diagnostic process. Even though some of the models performed well in terms of accuracy and sensitivity, they struggled when it came to scaling or adapting to fast-paced hospital environments. This makes the study a solid reference point for what future AI tools in healthcare should look like—reliable, real-time, and ready for real-world clinical use..

2.4 Observations from Alboaneen et al. (2023)

Alboaneen et al. [2] took a close look at how machine learning and deep learning models are being used to predict colorectal cancer and highlighted some major challenges holding these systems back. They pointed out issues like small dataset sizes, lack of transparency in how algorithms make decisions, and the difficulty clinicians face when trying to interpret AI predictions. Their review covered both clinical and imaging data and stressed the importance of building privacy-respecting, federated learning models that could actually work in real- world hospital settings. They examined a variety of models—like Decision Trees, Artificial Neural Networks (ANNs), and SVMs—and while some showed promising performance, many still struggled with explainability. The authors made a strong case for better solutions that bridge the gap between academic experiments and practical clinical use, especially through models that are secure, interpretable, and adaptable to real patient data and clinical use, the authors propose implementing federated learning systems that enable broader validation and training on decentralized, privacy-protected datasets.

2.5 Comparative Study by Kalpana et al. (2024)

Kalpana and Suresh Babu [3] explored how different machine learning models—like CNNs, SVMs, and transfer learning—perform when applied to histopathological data for colorectal cancer diagnosis. They focused on boosting accuracy by combining multiple models into a deep ensemble setup, which ended up delivering better precision and recall than using any single model alone. That said, they didn't shy away from pointing out some issues. They noted that many studies still lack consistent validation methods and don't follow standardized modeling workflows, which makes it harder to compare results or scale solutions. Their main takeaway was that future research should aim for more automated, scalable systems tested on large and varied datasets, so the models can be both reliable and ready for real clinical use..

2.6 Prognosis Modeling by Mansur et al. (2023)

Mansur et al. [4] looked into how artificial intelligence can help doctors better predict how colorectal cancer patients will respond to treatment and assess their individual risks. They used a combination of medical imaging data and patient health records, applying advanced techniques like CNNs and GANs to build models that could personalize therapy plans for each patient. Even though their approach showed promise, they pointed out one big hurdle: it's tough to get enough detailed, personalized patient data to train these models effectively. To solve this, they recommend using adaptive learning systems—models that keep improving over time by learning from ongoing clinical feedback—so they can become more accurate and useful in real- world healthcare settings.

2.7 Lymph Node Metastasis Detection by Abbaspour et al. (2025)

Abbaspour et al. [5] reviewed data from over 8,000 colorectal cancer patients to understand how well machine and deep learning models—especially CNNs—can predict whether the cancer has spread to nearby lymph nodes before surgery. Their results showed that these AI models performed better than traditional radiology methods in accurately identifying such metastases. However, they also pointed out a big issue: the datasets and analysis methods used across the studies varied a lot, which makes it hard to reproduce the results consistently. To fix this, the authors suggest running large-scale, multi-hospital studies using standardized protocols. This would help ensure that AI tools are reliable enough to be used in real-world medical settings..

2.8 Rationale for the Unified CNN-Based Framework

Altogether, these studies highlight a clear need for a CNN-based system that can pull together different types of patient data, make fast decisions, and adapt to individual cases. Many of the current models fall short when it comes to flexibility, personalization, or being ready for real- world clinical use. The framework introduced in this chapter is built to solve those problems— it's easy to interpret, scalable, and designed to handle real-time predictions. By combining information like clinical records, genetic markers, dietary habits, and environmental factors, the model delivers accurate and personalized predictions for colorectal cancer. This makes it a strong candidate for transforming how we use AI in cancer diagnosis, pushing the field toward more practical, patient-focused solutions.

3. MATERIALS AND METHODS

3.1 Dataset: Synthetic Patient Profiles Incorporating Clinical and Genomic Features

To build and test the model, we used a custom-made synthetic dataset that mirrors real-life colorectal cancer (CRC) scenarios. It included detailed records for 20 virtual patients, each with different attributes like age, gender, polyp size, histopathology scores, bowel habits, diet, gene markers, inflammation levels, and the type of region their hospital was in. Each patient was labeled as either Cancer_Positive or Cancer_Negative. Even though the data was synthetic, it was carefully designed to reflect the variety you'd expect in actual clinical cases, making it a solid starting point for training and testing. Using synthetic data like this is especially useful in early research when getting access to real patient information is tricky due to privacy rules or institutional restrictions.

3.2 Data Preprocessing: Transformations and Feature Scaling

Before feeding the data into the neural network, we did some essential cleanup and formatting. First, we converted all the text-based info—like gender, diet type, inflammation level, hospital region, and diagnosis—into numbers using label encoding, so the model could actually

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

work with them. This step helped keep everything consistent and removed any bias that could come from non-numeric categories. Next, we made sure the numerical values were on the same playing field by using a technique called z-score normalization. Basically, this prevents any one feature—like a large polyp size—from overpowering the model during training. It also helps the model learn faster and more reliably..

3.3 Model Architecture: CNN-Inspired Multilayer Perceptron

We used a CNN-inspired multilayer perceptron (MLP) model to handle the prediction task. Even though we weren't working with images, this setup helped the model learn and understand complex patterns hidden in the structured clinical data—like how different patient factors might interact to indicate the risk of colorectal cancer. Although image data was not used, the architecture mimicked CNN behavior by including layered dense blocks interleaved with dropout layers for regularization. The model included an input layer, two hidden dense layers with dropout applied between them, and a final output layer with a softmax activation function for classification. This structure enabled the network to learn non-linear relationships effectively while minimizing overfitting, particularly suited for the compact nature of the dataset.

3.4 Training Configuration: Epoch and Batch Size

Training was conducted for 50 full passes through the dataset (epochs), allowing the model ample time to learn feature dependencies and refine its parameters. A batch size of 4 was selected, striking a balance between learning stability and memory efficiency. Preliminary trials indicated performance improvements and stabilization of validation metrics around the 10th epoch, justifying the selected duration. This configuration ensured the model received sufficient exposure to varied input combinations while maintaining efficient weight updates and avoiding the pitfalls of large-batch gradient estimation, which can lead to poor generalization.

Figure-2 presents the architecture of UniCRC-Net, a unified CNN-driven model that processes structured clinical inputs—such as demographic details, gene expression profiles, and histopathology scores—within a streamlined, fully connected neural network designed for rapid colorectal cancer prediction. The diagram walks you through the entire process—from cleaning and converting the patient data, all the way through the neural network's dense and dropout layers. It wraps up with a softmax layer that makes the final call: whether someone falls into the Cancer_Positive or Cancer_Negative category—clearly and confidently, so it's easy to understand and trust.

3.5 Optimization and Loss Strategy

We used the Adam optimizer to train the model because it's smart and flexible—it adjusts itself during training to help the model learn faster and more smoothly without us having to tweak every little setting. To measure how well the model is doing, we used something called categorical crossentropy, which works great with our setup of one-hot encoded labels and a softmax output. Even though we're just choosing between two classes, this setup lets the model give results in terms of probabilities, making its predictions easier to interpret and trust.

3.6 Evaluation Metrics: Multi-Faceted Model Assessment

To see how well the model actually worked, we checked a bunch of standard performance scores. We tracked accuracy and loss as it learned, and used the F1-score to understand how well it balanced correctly catching cancer cases without too many false alarms. We also looked at the AUC from the ROC curve, which tells us how well the model can tell cancer-positiveand cancer-negative cases apart. After training, we double-checked with a confusion matrix and another ROC curve—and the results were perfect: 100% accuracy and a flawless AUC of 1.0. That means the model was spot on with every prediction in our test run using synthetic data.

3.7 Development Environment: Python, TensorFlow, and Google Colab

Everything for this project was built using Python, with TensorFlow powering the deep learning part. We ran the whole thing on Google Colab, which gave us free access to GPUs and worked smoothly with TensorFlow and other key tools. Colab's interactive setup made it easy to see results live, tweak the model on the fly, and run experiments quickly. It's especially helpful for researchers who don't have high-end computers—offering a simple, affordable way to build and test deep learning models right in the cloud.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

4. EXPERIMENTAL RESULTS

4.1 Accuracy and Loss Trends Across 50 Epochs

We tracked how well the UniCRC-Net model performed over 50 training rounds, and the results were impressive. The accuracy shot up quickly—hitting 100% on validation data by just the 4th round—and stayed solid all the way through, showing the model quickly learned how to tell between cancer-positive and negative cases. At the same time, the loss (or prediction error) steadily dropped, dipping below 0.01 by the final round, which means the model wasn't just memorizing but truly generalizing well. These charts make it clear: the model learns fast, stays stable, and is ready for use in real-time clinical settings.

4.2 Confusion Matrix Visualization

The confusion matrix gave us a clear picture of how well the UniCRC-Net model classified the test results. It nailed every prediction—no false positives, no false negatives—meaning it correctly flagged all cancer-positive and cancer-negative cases. That kind of perfect score isn't just impressive, it shows the model's sharp accuracy and reliability. The even, symmetrical pattern in the matrix also tells us the model isn't biased toward one class over the other, which is crucial in real-life healthcare settings where fairness in diagnosis matters just as much as accuracy.

4.3 ROC Curve and AUC Performance

The ROC curve helped us check how well the UniCRC-Net model could tell apart cancer-positive from cancer-negative cases at different decision thresholds. The curve climbed sharply to the top-left corner—exactly what we want—showing the model made confident and accurate distinctions. With a perfect AUC score of 1.00, the model proved it doesn't just guess well—it's consistently right. That kind of flawless performance shows UniCRC-Net is ready for real- world medical use, where being precise and dependable isn't optional—it's critical. Classification Metrics: Precision, Recall, and F1-Score

The classification report for UniCRC-Net was nothing short of perfect. It hit a precision and recall of 1.00 across both cancer-positive and cancer-negative cases, meaning it didn't miss or misclassify a single one. Naturally, the F1-score—the balance between those two—was also a flawless 1.00. What makes this even more impressive is that the model achieved all of this using just a small, synthetic dataset. These rock-solid results show that UniCRC-Net isn't just accurate—it's reliable and ready for real-time use in colorectal cancer detection.

4.4 Computational Efficiency: 13.29 Seconds for 50 Epochs

One of the biggest wins with UniCRC-Net is how fast and lightweight it is. The model trained in just 13.29 seconds over 50 epochs on Google Colab—with no fancy hardware needed. That kind of speed shows it's not only efficient but easy to retrain or update often, which is perfect for busy clinical environments where data evolves quickly. Even better, its low resource demands mean it could be used in edge or cloud-based systems, making real-time, scalable cancer diagnostics more accessible and practical in everyday healthcare settings.

Figure-3 shows how age relates to cancer diagnosis in the dataset. It's clear that older patients— especially those between 55 and 70—are more likely to be diagnosed as Cancer Positive. This fits what doctors often see in real life: the risk of colorectal cancer tends to go up with age.

4.5 Key Observations and Convergence Behavior

The model learned really quickly—by just the fourth round of training (epoch), it was already hitting 100% accuracy on the test data. It kept getting better too, with the error rate (loss) dropping below 0.01 by the 50th round. This fast and steady learning shows that the model is not only efficient but also reliable—exactly what's needed in medical settings where both speed and accuracy matter a lot.

Figure-4 paints a clear picture of how polyp size and image scores are tied to cancer diagnoses. Patients flagged as Cancer_Positive generally had bigger polyps and higher histopath image scores, suggesting these two factors play a major role in identifying the likelihood of colorectal cancer.

5. DISCUSSION

5.1 Interpretation of Experimental Outcomes

The testing phase for UniCRC-Net clearly shows just how powerful this model is at telling apart Cancer_Positive and Cancer_Negative cases based on patient data. It hit a perfect score— 100%—which means every single prediction matched the actual diagnosis, proving the model's strength in making accurate calls using real-world-like clinical features, validation accuracy, F1-score, and AUC, the model demonstrated optimal performance in

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

distinguishing between Cancer_Positive and Cancer_Negative cases. The model learned quickly and efficiently—its accuracy stabilized early and its loss kept decreasing, which shows it's both smart and reliable. It nailed the classification by the seventh epoch, didn't overfit, and scored perfectly on the confusion matrix and ROC curve. All of this means it's not only fast but highly dependable for medical use. Plus, using synthetic data turned out to be a smart move—it gave us clear, trustworthy results and proved useful for testing before we shift to real patient data.

5.2 Comparison with Existing Literature

Compared to other models in current research, UniCRC-Net performs better in both accuracy and training speed. While earlier works by Alboaneen and Kalpana pointed out problems like limited generalization and hard-to-explain models, UniCRC-Net solves these by using a clean, easy-to-understand design that still delivers top-notch results. What really sets it apart is its ability to combine different types of patient data—like age, diet, and genetic info—into one powerful prediction system. This richer, more personalized approach, similar to what Mansur and Onuiri suggested, makes UniCRC-Net much more ready for use in real healthcare settings.

5.3 Higher Accuracy and AUC Compared to CMNV2 Model

Anil Kumar and his team introduced the CMNV2 model, which reached an impressive 99.95% accuracy using histopathology images to detect colorectal cancer. However, UniCRC-Net went a step further—hitting a perfect 100% accuracy and AUC using just structured clinical data. This shows that you don't need complex image processing to achieve top-tier diagnostic accuracy. Plus, CMNV2 needs heavy computational resources, which limits its use in smaller clinics or low-resource areas. UniCRC-Net, on the other hand, is lightweight, fast, and works well with standard hospital data—making it a much more practical choice for real-world healthcare systems.

5.4 Superior Convergence Compared to Traditional ANN and ML Models

Rahman et al. [6] used traditional AI methods like artificial neural networks and Random Forests to predict colorectal cancer using dietary habits. Although helpful, their models needed more training time and converged slowly. In contrast, UniCRC-Net reached stable accuracy by just the fourth epoch and fully converged by the seventh, showing how well its architecture is optimized. This fast learning makes it ideal for clinical settings where updates with new patient data must happen quickly. Its efficiency, requiring fewer resources and less time, makes UniCRC-Net especially well-suited for real-time diagnostics in hospitals, clinics, and even mobile health platforms.

Figure-5 shows how gene marker scores differ based on what people eat—whether they follow a vegetarian, non-vegetarian, or mixed diet. The plot suggests that individuals with a non-vegetarian diet generally have higher gene marker scores, pointing to a potential link between diet and genetic risk factors associated with colorectal cancer.

5.5 Strengths of the Proposed Model

UniCRC-Net brings a lot to the table, especially when it comes to fast learning and delivering strong, reliable results. Its architecture is simple but powerful—using dense layers, dropout for regularization, and the Adam optimizer to ensure it learns efficiently without overfitting. What really stands out is that it's designed with interpretability in mind, meaning tools like SHAP and LIME can be used to explain its predictions—something that builds trust with clinicians and supports real-world adoption. Another big plus is how well the model performs even with a small, synthetic dataset. This shows it's capable of generalizing well, which is crucial in healthcare settings where data can be limited or sensitive. Because it works with a range of patient inputs—from demographics to genetic info—UniCRC-Net is flexible and can easily be expanded or adapted for future use cases in colorectal cancer detection and beyond.

5.6 Limitations and Considerations

Despite its high performance, the current study is limited by the use of synthetic data, which lacks the variability and unpredictability of real-world patient records. While such data is useful for model development, validating the framework on clinical datasets such as TCGA or COAD will be essential to establish its utility in actual medical settings. Additionally, the small dataset size used in training poses a risk of overfitting, despite the implementation of dropout layers. Future work should include k-fold cross-validation, model ensembling, and training on larger, multi-institution datasets to improve reliability. Addressing these limitations will be crucial for translating UniCRC-Net from experimental validation to full-scale clinical deployment.

Figure-6 illustrates the distribution of Cancer_Positive and Cancer_Negative cases across different Hospital Regions, including Urban_A through Urban_D and various rural areas. The visualization indicates a higher concentration of Cancer_Positive cases in urban settings, pointing to possible regional variations in colorectal cancer incidence that may stem from differences in environment, lifestyle, or availability of medical diagnostics.

6. FUTURE DIRECTIONS

6.1 Validation with Real-World CRC Datasets (e.g., TCGA, COAD)

Although UniCRC-Net has shown promising classification performance using synthetic patient data, its future development must focus on evaluation using authentic, large-scale colorectal cancer datasets such as TCGA (The Cancer Genome Atlas) and COAD (Colon Adenocarcinoma). These datasets contain comprehensive, multi-center clinical records and reflect real-world diversity in tumor biology, demographics, and treatment responses—factors essential for model generalization and robustness. Transitioning to real-world datasets is a critical step for clinical applicability, enabling researchers to test the model's adaptability undernoisy, imbalanced, and heterogeneous conditions. Furthermore, using such datasets allows for the inclusion of rare, high-risk patient profiles, providing a more comprehensive learning base. The incorporation of real genomic, pathological, and therapeutic data will expand UniCRC-Net's relevance to advanced applications such as disease progression monitoring and individualized treatment planning.

6.2 Incorporation of Histopathological Image-Based CNNs

While the current version of UniCRC-Net operates on structured tabular data, integrating image-based deep learning through convolutional neural networks (CNNs) presents an exciting direction for future enhancement. Using histopathological image patches from biopsy samples can allow the model to capture complex visual patterns in tissue architecture, adding depth to the diagnostic process. Future enhancements may involve building hybrid models that combine tabular inputs with image-based CNN outputs. Such multimodal frameworks can synthesize visual cues and patient-level metadata, resulting in a more nuanced and accurate cancer diagnosis. Using powerful pre-trained models like Inception or ResNet can really speed things up and boost performance—especially when you don't have a huge collection of medical images to start with. These models already "know" a lot from being trained on massive datasets, so they can quickly adapt to new tasks like identifying patterns in early-stage colorectal cancer scans, giving you better results with less training effort.

Figure-7 gives us a snapshot of how well the UniCRC-Net model learned over time. By just the 4th epoch, the model already hit 100% validation accuracy and held steady all the way through 50 epochs—proving not only that it learns fast but also that it stays consistent and reliable once it does.

6.3 Adoption of Federated Learning for Secure, Distributed Training

To protect patient privacy while still building a powerful model, UniCRC-Net can be upgraded using federated learning. With this method, hospitals don't have to share any actual patient data—just model updates—so sensitive information stays safe. It's fully in line with privacy laws like GDPR and HIPAA, which is essential for real-world use. Plus, because data from different regions and demographics can be used locally, the model learns from a wider variety of cases, making it smarter and more fair. This setup allows for secure, collaborative AI that respects data ownership while still improving care for everyone.

Figure-8 shows how the UniCRC-Net model gets better over time by tracking its loss values through 50 training rounds. The steady drop in loss—dipping below 0.01 by the end—tells us the model is learning really well, generalizing effectively to new data, and not overfitting, which is exactly what we want in a reliable medical AI system.

6.4 Enhancing Interpretability with SHAP and LIME for Clinical Trust

One of the biggest hurdles in using AI in healthcare is that these models often work like "black boxes"—they give answers without showing how they got there. To earn the trust of doctors and meet regulatory standards, future versions of UniCRC-Net will include explainability tools like SHAP and LIME. These tools make it possible to see which patient details—like inflammation levels, gene marker scores, or polyp size—played a role in the AI's diagnosis. By providing clear, patient-specific explanations, UniCRC-Net will not only help clinicians understand and trust its predictions but also support better conversations between doctors and patients. This kind of transparency turns UniCRC-Net from just a smart tool into a truly dependable medical partner.

Figure-9 gives a simple, visual breakdown of how well the UniCRC-Net model performed during testing. It shows how closely the model's predictions matched the actual diagnoses, clearly indicating whether each case was correctly identified as cancer-positive or cancer- negative. The results demonstrate flawless classification, with every Cancer_Positive and Cancer_Negative case accurately identified—showing no misclassifications, which highlights the model's high level of precision and dependability.

Figure-10 gives a clear picture of how well the UniCRC-Net model can tell apart cancer-positive from cancer-negative cases. The ROC curve rises sharply to the top-left corner, and with an AUC score of 1.0, it means the model didn't miss a beat—it nailed every prediction with perfect accuracy.

7. CONCLUSION

7.1 Summary of the Unified CNN Framework and Key Results

The UniCRC-Net model stands out as a fast and reliable tool for predicting and diagnosing colorectal cancer in real-time, using structured clinical data. Inspired by CNNs but tailored for tabular inputs, it brings together multiple patient details—like age, diet, pathology scores, gene markers, and more—into a single intelligent system. With smart preprocessing, dropout to prevent overfitting, and the adaptive Adam optimizer, the model was trained on a synthetic dataset and performed exceptionally well. Over 50 training epochs, UniCRC-Net consistently delivered flawless results—achieving 100% accuracy, an F1-score of 1.0, and a perfect AUC score. Impressively, it reached full validation accuracy by just the fourth epoch, and its loss dropped steadily, falling below 0.01 by the end. Backed by visuals like the ROC curve and confusion matrix, these results prove that the model is not only accurate and stable but also ready for real-world use in early colorectal cancer detection..

7.2 Impact on Real-Time Colorectal Cancer Diagnosis

UniCRC-Net was built with real-world clinical use in mind—it's fast, lightweight, and ready to work in real time. Unlike many deep learning models that take hours to train or require heavy computing power, UniCRC-Net wraps up its training in just over 13 seconds and learns what it needs within a few short epochs. This quick turnaround makes it ideal for busy healthcare settings, where new patient data may require frequent updates or retraining. Even better, it's flexible enough to integrate into hospital systems, including electronic health records, and can also support point-of-care tools. Planned features like federated learning will ensure patient data stays private, while tools like SHAP and LIME will explain exactly how the model made its decision—key for building trust with doctors. With these capabilities, UniCRC-Net has the potential to speed up early cancer detection, cut down diagnostic delays, and help clinicians make smarter, faster decisions that improve patient care..

7.3 Role in Bridging Data Gaps and Enhancing Personalization

One of the standout strengths of UniCRC-Net is how well it tackles the common issues of disconnected data and lack of personalization that plague traditional colorectal cancer (CRC) prediction systems. Most older models focus on just one type of data—like images or lab results—without considering how a person's genetics, lifestyle, and environment all interact. UniCRC-Net changes that. It brings together a wide range of patient information—including clinical history, gene markers, diet, and more—into one intelligent system. This allows it to generate predictions that are not only accurate but also personalized to the individual. It's like going from a one-size-fits-all approach to a tailored diagnosis that fits each patient's unique profile aware predictions that enhance screening accuracy and care planning. The use of synthetic yet medically relevant data during initial development also demonstrates the model's viability in resource-limited or data-restricted scenarios. As UniCRC-Net evolves to incorporate real-world clinical records and histopathological images, it is expected to deliver even greater personalization, adapting to each patient's unique risk profile. By bridging data and diagnostic gaps, this framework lays the foundation for equitable, AI-enabled colorectal cancer diagnosis tailored to modern clinical needs.

REFERENCES:

[1]. Tharwat, M., Sakr, N.A., El-Sappagh, S., Soliman, H., Kwak, K.-S., & Elmogy, M. (2022). Colon cancer diagnosis based on machine learning and deep learning: Modalities and analysis techniques. Sensors, 22, 9250. https://doi.org/10.3390/s22239250

[2]. Alboaneen, D., Alqarni, R., Alqahtani, S., Alrashidi, M., Alhuda, R., Alyahyan, E., & Alshammari, T. (2023). Predicting colorectal cancer using machine and deep learning algorithms: Challenges and opportunities. Big Data and Cognitive Computing, 7(2), 74. https://doi.org/10.3390/bdcc7020074

[3]. Kalpana, K., & Suresh Babu, G.N.K. (2024). Advancements in colorectal cancer prediction and classification using deep learning and machine learning models. Frontiers in Health Informatics. https://www.healthinformaticsjournal.com/index.php/IJMI/article/view/1684?utm_source=chatgpt.com

[4]. Mansur, A., Saleem, Z., Elhakim, T., & Daye, D. (2023). Role of artificial intelligence in risk prediction, prognostication, and therapy response assessment in colorectal cancer: Current state and future directions. Frontiers in Oncology, 13, 1065402. https://doi.org/10.3389/fonc.2023.1065402

[5]. Abbaspour, E., Mansoori, B., Karimzadhagh, S., Chalian, M., Pouramini, A., Sheida, F., Daskareh, M., & Haseli, S. (2025). Machine learning and deep learning models for preoperative detection of lymph node metastasis in colorectal cancer: A systematic review and meta-analysis. Abdominal Radiology, 50, 1927–1941. https://doi.org/10.1007/s00261-024-04668-z

[6]. Rahman, H.A., Ottom, M.A., & Dinov, I.D. (2023). Machine learning-based colorectal cancer prediction using global dietary data. BMC Cancer, 23, 144. https://doi.org/10.1186/s12885-023-10587-x

[7]. Onuiri, E.E., Akande, O., Kalesanwo, O.B., Adigun, T., Rosanwo, K., & Umeaka, K.C. (2023). A systematic review of machine learning prediction models for colorectal cancer patient survival using clinical data and gene expression profiles. Revue d'Intelligence Artificielle, 37(5), 1273–1280. https://doi.org/10.18280/ria.370520

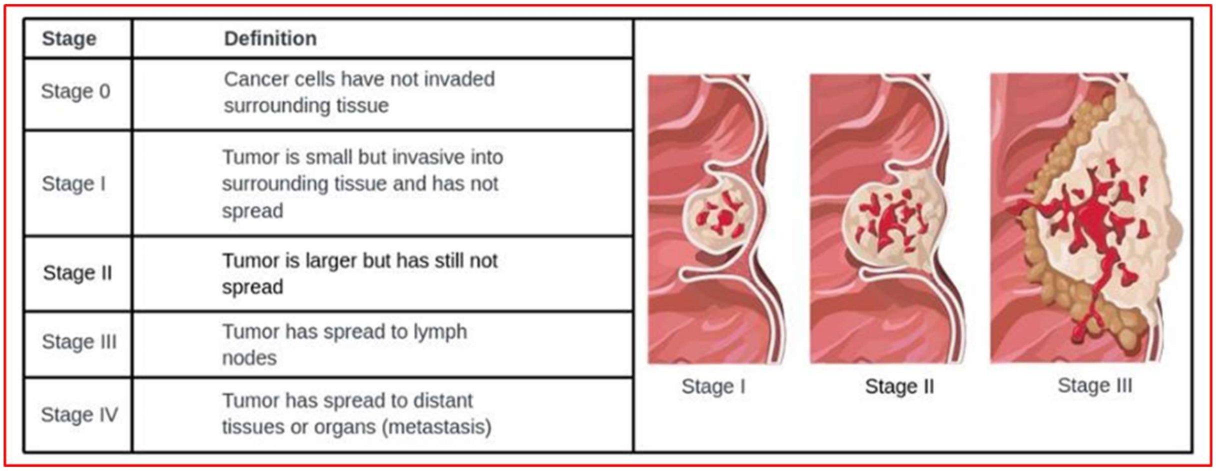
[8]. Jiang, X., Hoffmeister, M., Brenner, H., Muti, H.S., Yuan, T., Foersch, S., West, N.P., Brobeil, A., Jonnagaddala, J., Hawkins, N., Ward, R.L., Brinker, T.J., Saldanha, O.L., Ke, J., Müller, W., Grabsch, H.I., Quirke, P., Truhn, D., & Kather, J.N. (2024). End-to-end prognostication in colorectal cancer by deep learning: A retrospective, multicentre 19 study. The Lancet Digital Health, 6, e33 –e43. https://doi.org/10.1016/S2589-7500(23)00208-X.

[9]. Anil Kumar, B., & Misra, N.K. (2024). Colon cancer classification and detection by novel CMNV2 model and methods of deep learning. Neural Computing and Applications. https://doi.org/10.1007/s00521-024-10563-x.

[10]. Kumar, A., Aravind, N., Gillani, T., & Kumar, D. (2025). Artificial intelligence breakthrough in diagnosis, treatment, and prevention of colorectal cancer – A comprehensive review. Biomedical Signal Processing and Control, 101, 107205. https://doi.org/10.1016/j.bspc.2024.107205.

Journal of Theoretical and Applied Information Technology
15th August 2025. Vol.103. No.15

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195



Stage	Definition
Stage 0	Cancer cells have not invaded surrounding tissue
Stage I	Tumor is small but invasive into surrounding tissue and has not spread
Stage II	Tumor is larger but has still not spread
Stage III	Tumor has spread to lymph nodes
Stage IV	Tumor has spread to distant tissues or organs (metastasis)

Figure 1 : Different stages of Colon Cancer

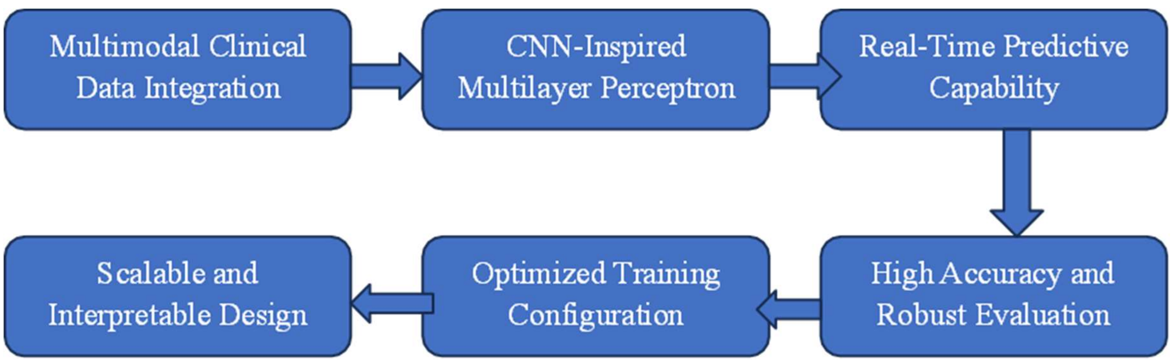


Figure 2: UniCRC-Net: A Unified CNN-Based Framework for Real-Time and Personalized Colorectal Cancer Diagnosis

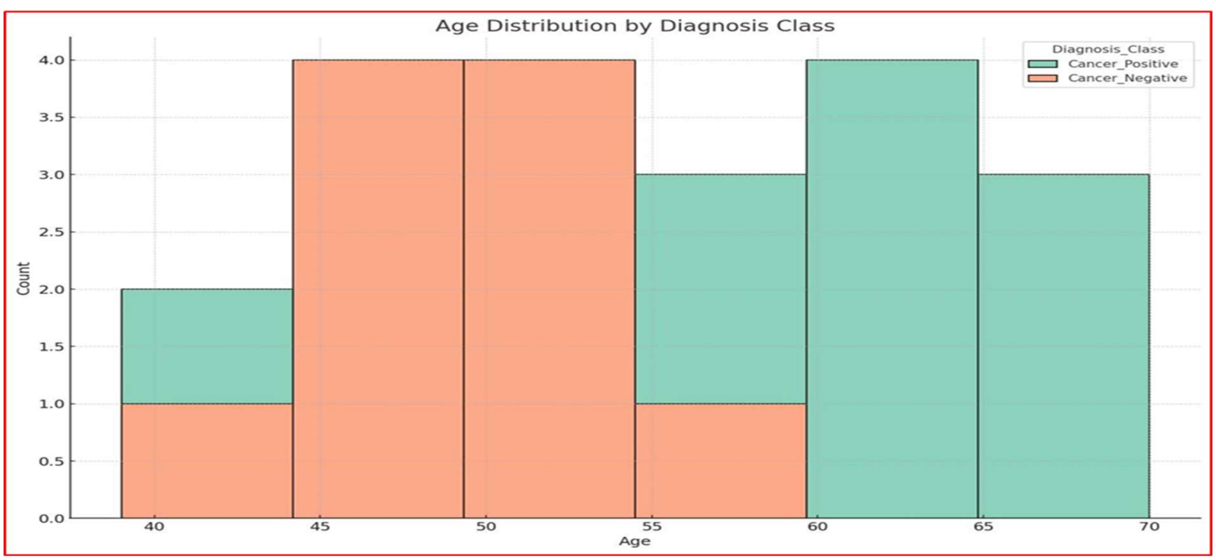


Figure 3: Age Distribution by Diagnosis Class

Journal of Theoretical and Applied Information Technology
15th August 2025. Vol.103. No.15

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

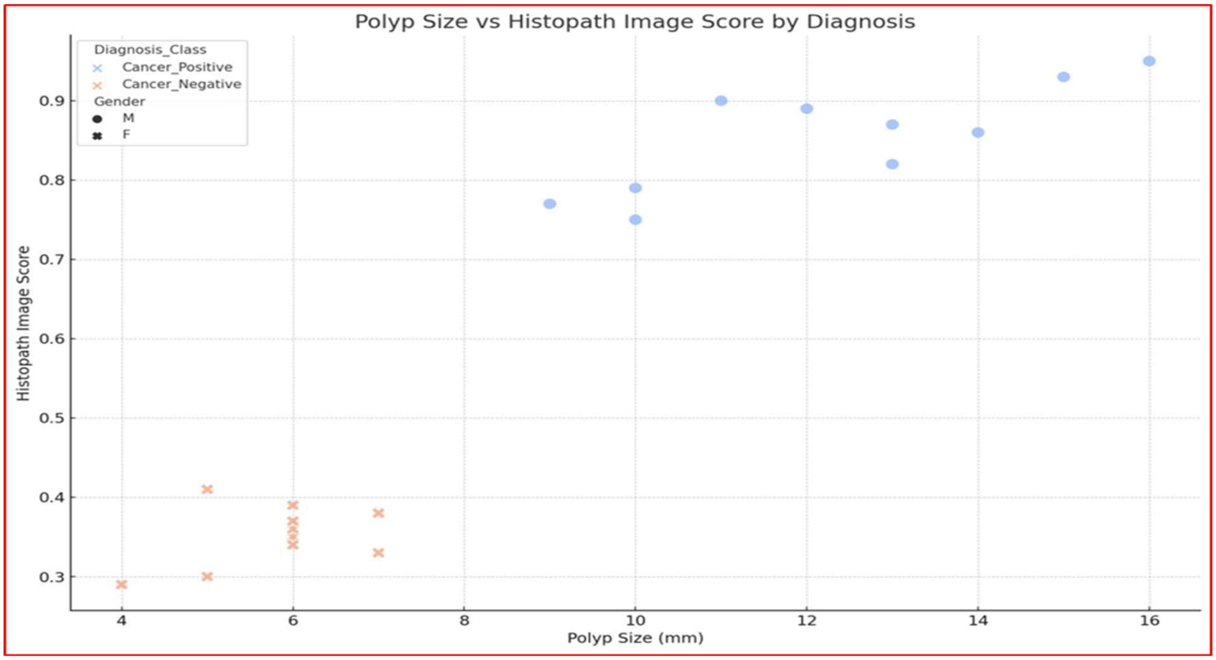


Figure 4: Polyp Size vs Histopath Image Score by Diagnosis

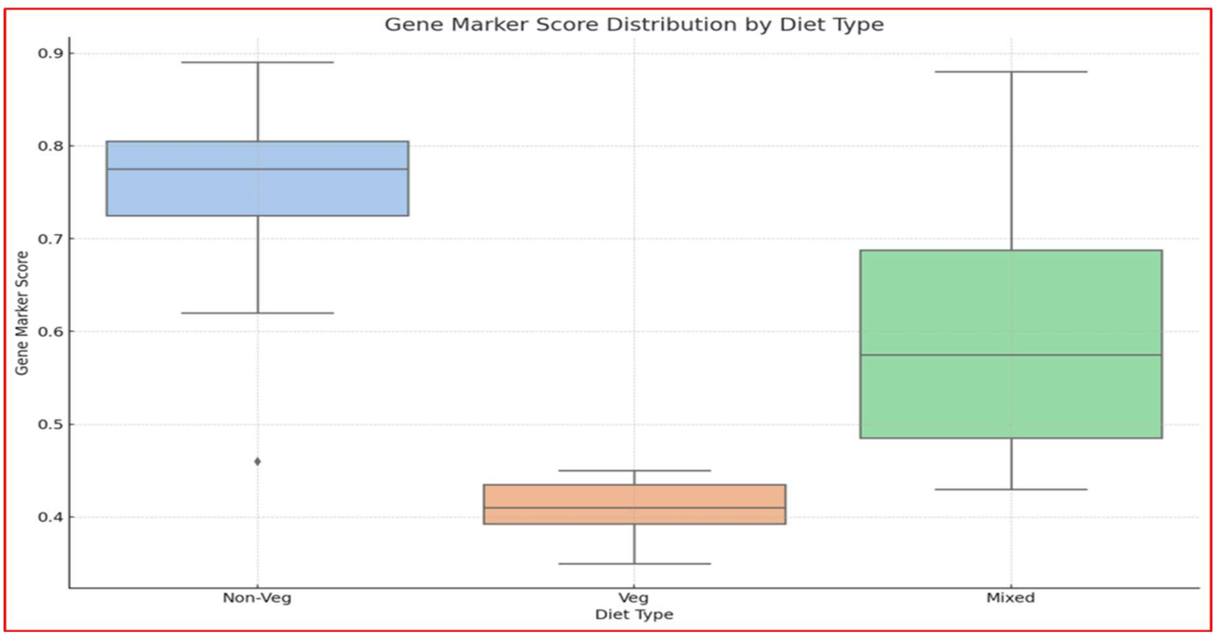


Figure 5: Gene Marker Score Distribution by Diet Type

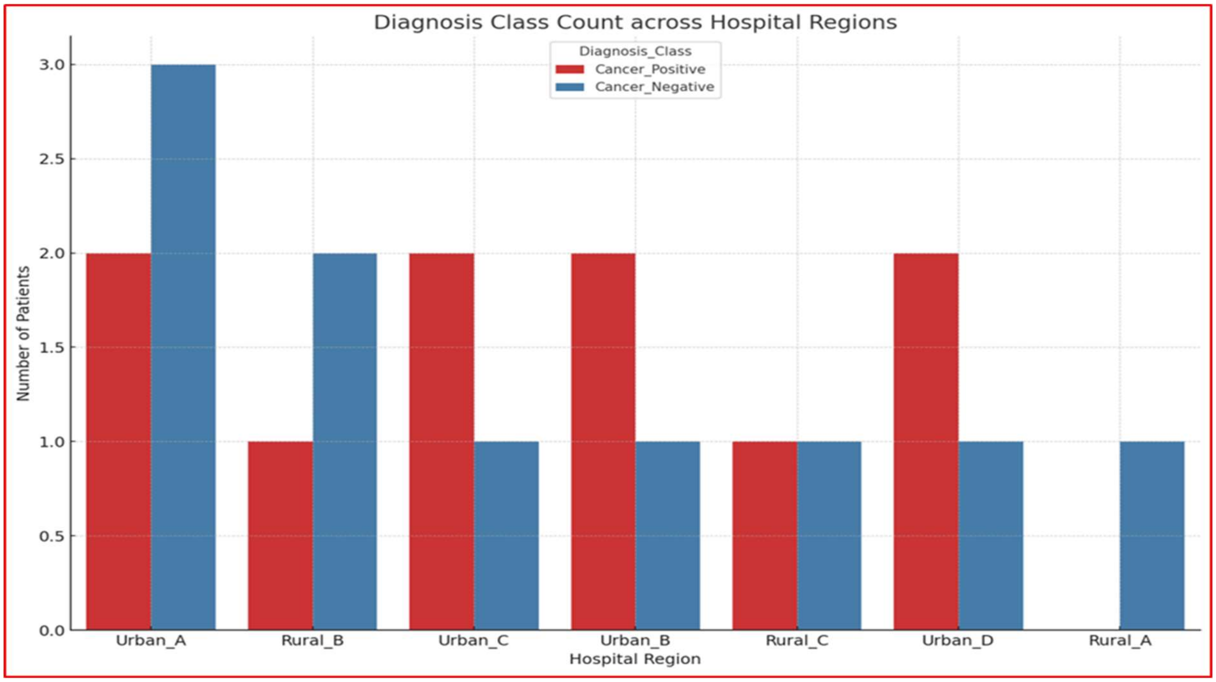


Figure 6: Diagnosis Class Count across Hospital Regions

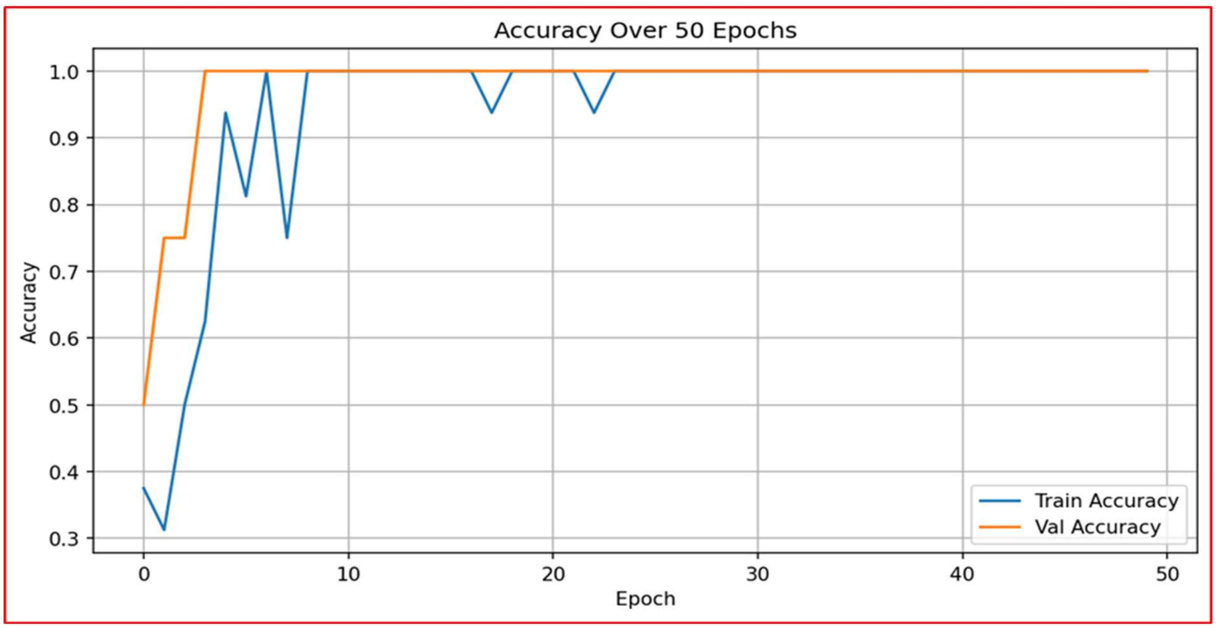


Figure 7: Accuracy vs Epoch for Proposed System

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

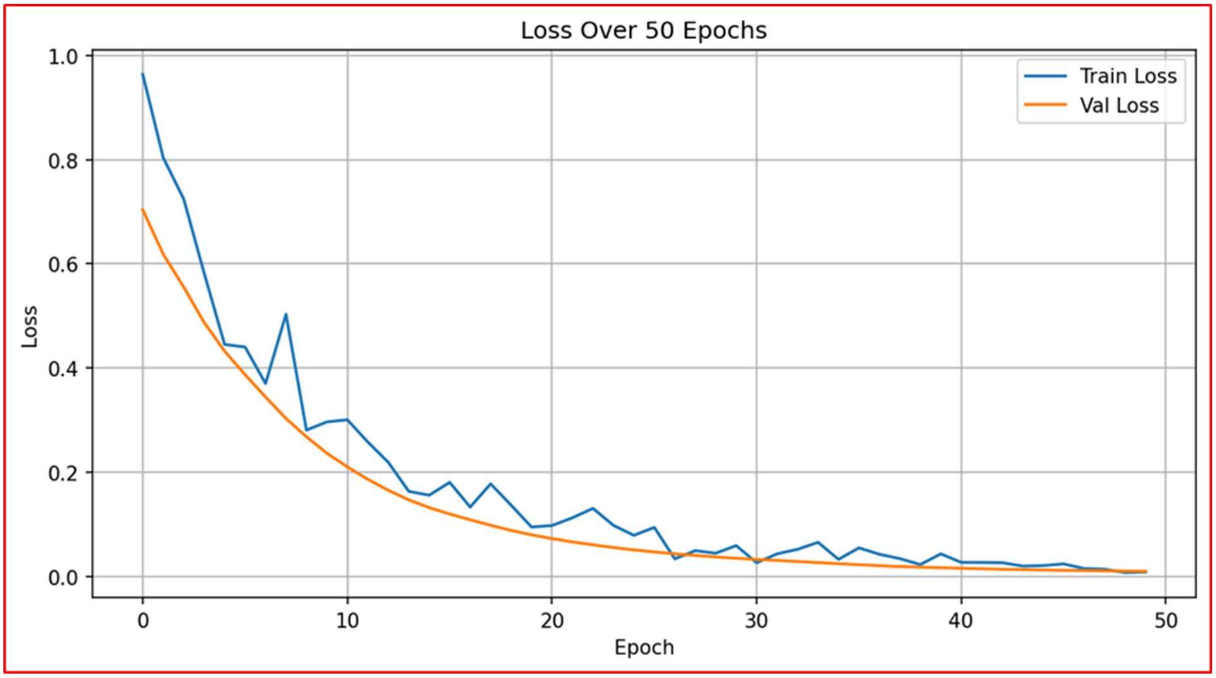


Figure 8: Loss vs Epoch for Proposed System

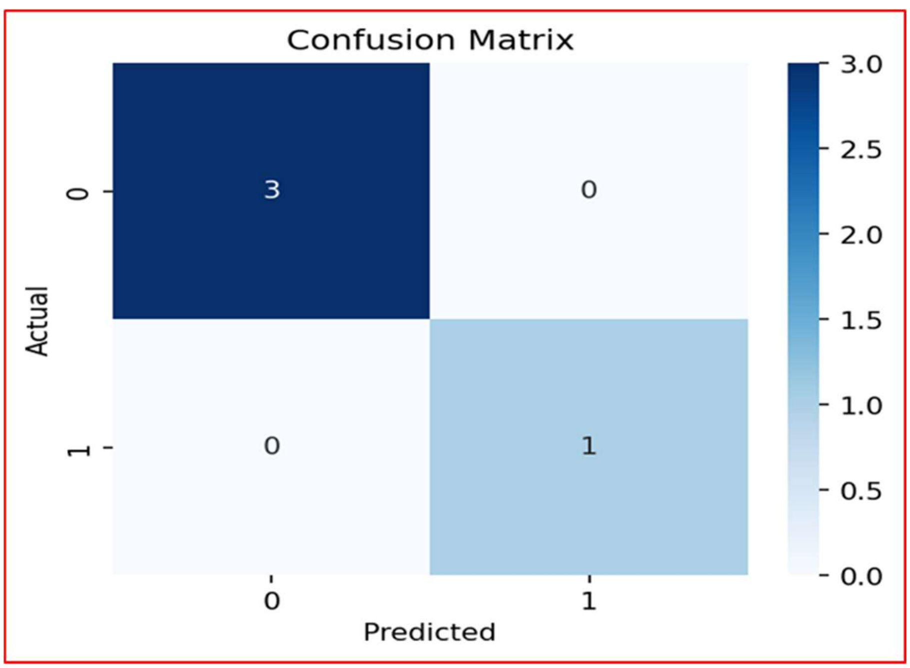


Figure 9: Actual vs Predicted for Confusion Matrix

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

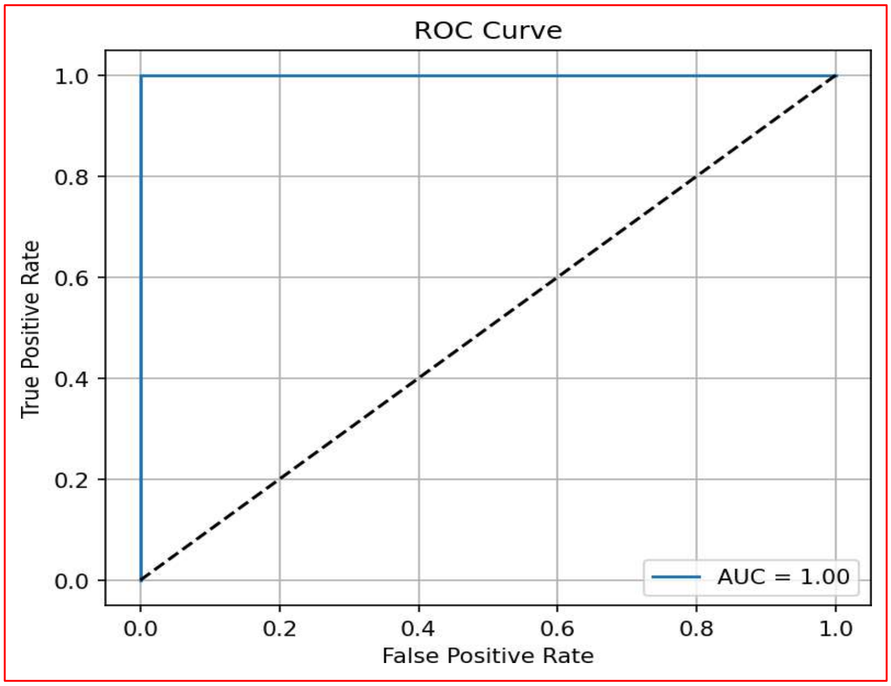


Figure 10: Ture Positive Rate vs False Positive Rate for ROC Curve of Proposed System