Review of Communication Research
ISSN: 2255-4165
https://rcommunicationr.org/index.php/rcr/

From Data to Dialogue: A Communication-Centered Perspective on AI-Based Multimodal Prediction of Neonatal Jaundice

Junkai Li [1], Bo Sun [1], Mohd Rizon Mohamad Juhari [1*], Tiang Sew Sun [1*]

[1] Faculty of Engineering, Technology and Built Environment, UCSI University, Kuala Lumpur 56000, Malaysia

* **Corresponding Author:** mohdrizon@ucsiuniversity.edu.my
tiangss@ucsiuniversity.edu.my

ARTICLE INFO

Received: 13 Jul 2025
Accepted: 9 August 2025

ABSTRACT

Neonatal jaundice is a common issue in the newborn period, and timely, accurate monitoring and prediction are crucial to prevent severe complications (e.g. kernicterus). However, traditional diagnosis relies on invasive serum bilirubin tests or subjective judgment, which limits applicability in developing countries and home settings. In recent years, machine learning techniques have been introduced for non-invasive neonatal jaundice prediction: convolutional neural networks (CNNs) have been used for analyzing skin images of jaundice, and long short-term memory (LSTM) networks for time series prediction of bilirubin changes, achieving some success. Yet, CNNs and LSTMs have limitations such as local receptive fields and difficulty capturing long-term dependencies. This paper comprehensively reviews and reconstructs the machine learning framework for neonatal jaundice prediction by introducing the Transformer model as the core. We use a Vision Transformer (ViT) instead of CNN for processing skin images, and a time-series Transformer in place of LSTM for modeling dynamic data like transcutaneous bilirubin readings, and we fuse multimodal information to improve prediction accuracy. We highlight the potential of this Transformer-based framework in the health communication domain: for example, using attention heatmap visualization to improve model interpretability, and integrating with mobile health (mHealth) applications for remote monitoring and interaction, thereby increasing public awareness and trust in new technology. We detail the methodological architecture, demonstrate its promising application in early prediction of neonatal jaundice, discuss current challenges, and envision future research directions. By leveraging a Transformer-centric multimodal deep learning framework, neonatal jaundice prediction could achieve higher accuracy and interpretability, better serving clinical decision-making and family healthcare.

Keywords: Artificial Intelligence (AI), Multimodal Prediction, Neonatal Jaundice, Healthcare Communication.

INTRODUCTION

Neonatal jaundice refers to the yellowing of the skin and sclera of newborns due to elevated bilirubin in the blood. Most newborns develop physiological jaundice 2–5 days after birth, which usually resolves on its own within a week or two. However, if bilirubin levels rise abnormally high and are not promptly managed, pathological jaundice may develop, and severe cases can lead to bilirubin encephalopathy (kernicterus) and irreversible neurological damage. Traditional diagnosis primarily relies on visual assessment combined with serum bilirubin measurements, but repeated blood draws cause pain to neonates and are impractical for home monitoring. Moreover, in resource-limited primary care settings, quick and affordable testing methods are often lacking. Therefore, there is an urgent need for a precise and non-invasive jaundice monitoring technology to enable early identification of high-risk infants and timely intervention, improving neonatal health management.

With advances in artificial intelligence, data-driven machine learning methods offer new approaches for monitoring and predicting neonatal jaundice. Early research has explored using machine learning models to predict high bilirubin risk from neonatal clinical indicators and skin images. For example, traditional algorithms

like support vector machines (SVMs) and decision trees have been used to build jaundice prediction models based on a few clinical parameters, achieving a certain degree of improved prediction accuracy (Huang et al., 2018). In recent years, deep learning has become a research hotspot. CNNs in particular have made remarkable achievements in medical image analysis and have been applied to neonatal skin photo jaundice detection with good results: these models can extract features from skin color that correlate with bilirubin levels, enabling automated assessment of jaundice severity (Shin et al., 2016). Meanwhile, recurrent neural networks like LSTM have been used to model time-series data such as continuous transcutaneous bilirubin (TcB) readings, capturing trends in bilirubin changes to provide a basis for dynamic prediction. These studies show that data-driven algorithms have the potential to assist or even partly replace traditional invasive testing, enabling early warning and real-time monitoring of jaundice risk (Taylor et al., 2017).

However, existing methods face challenges: CNNs focus on local feature extraction and may overlook global skin color distribution patterns; LSTMs can retain some long-term dependencies, but when sequences are very long or multi-variable inputs are involved, they are difficult to train and inherently lack a global attention mechanism. Therefore, the key to further improving neonatal jaundice prediction accuracy is how to fuse image and time-series multimodal data, fully mine their correlations, and enhance the model's ability to capture long-term dependencies and global patterns. The rise of Transformer models offers an opportunity to address these issues. Transformers were first revolutionary in natural language processing with the "attention mechanism" (Vaswani et al., 2017), and their core multi-head self-attention mechanism can efficiently model global relationships among sequence elements. Inspired by this, researchers have begun applying Transformers to computer vision and time-series domains, giving rise to models like the Vision Transformer (ViT) for direct image analysis and various Transformer variants for time-series prediction (Dosovitskiy et al., 2021; Lim et al., 2021). These have achieved performance comparable to or even better than CNNs/RNNs on many tasks, especially when data is abundant, with Transformers demonstrating great potential due to their stronger modeling capacity.

Based on this, our work proposes a multimodal neonatal jaundice prediction framework centered on Transformers: using ViT to handle newborn skin images, using a time-series Transformer to analyze dynamic physiological data like TcB, and fusing the two for a more comprehensive and accurate prediction. In the Methods section we will describe the structure and working principle of this framework in detail, in the Results and Application section we will showcase its potential use, and in the Discussion and Challenges section we will analyze remaining issues to be solved – including how to leverage this new technology to promote health communication and interaction between doctors and the public.

METHODOLOGY

Our proposed method framework comprises a Vision Transformer sub-model for image processing, a time-series Transformer sub-model for sequential data processing, and a multimodal fusion module that combines these two modalities (see Figure 3 below). First, newborn skin images are captured via smartphone or digital camera and input to a ViT model to assess the degree of skin jaundice; simultaneously, time-series data such as transcutaneous bilirubin (TcB) readings or weight changes over days after birth are collected and input to a time-series Transformer model to capture trends in bilirubin dynamics. Finally, features from the image and time-series branches are fused to output an integrated prediction of high bilirubin risk. Below we explain the principles of each component.

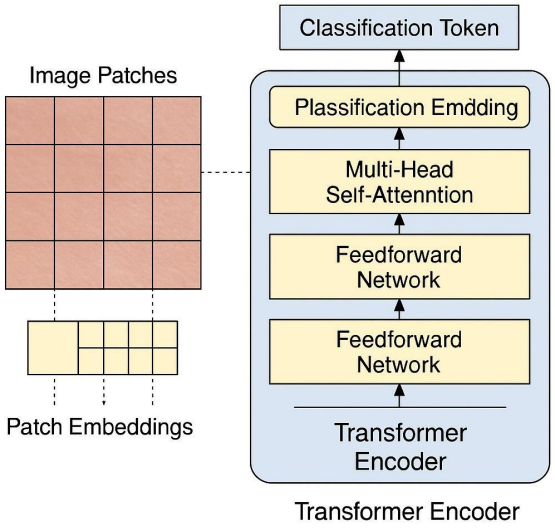


Figure 1. Vision Transformer (ViT) model structure diagram. ViT splits the input high-resolution skin image into a number of non-overlapping patches (for example, 16×16 pixels each). Each patch is flattened into a vector via a linear projection, and a positional encoding indicating its location in the original image is added to this patch embedding. These vectors, along with a learnable classification token, are fed as a sequence into a Transformer encoder. The Transformer encoder consists of multiple layers of alternating self-attention and feed-forward neural networks.

Scaled Dot-Product Attention in Vision Transformer (ViT)

In the Vision Transformer (ViT) model, each image is first divided into a series of fixed-size patches, which are linearly projected and combined with positional embeddings before being input into the Transformer encoder. Within each encoder layer, the self-attention mechanism computes the relationship between every patch and every other patch in the image. The mathematical form of single-head scaled dot-product attention is as follows:

$$\text{Attention}(Q, K, V) = \text{softmax}(QK^T / \sqrt{d_k}) \cdot V$$

Where:
- Q, K, and V represent the query, key, and value matrices respectively.
- d_k is the dimensionality of the key vectors.

This mechanism allows the model to compute weighted relationships between all image patches, thereby capturing global contextual information across the entire image.

Because the ViT does not impose the locality assumptions of convolution, it tends to overfit on small datasets; however, with large-scale pre-training or sufficient data augmentation, it performs very well (Dosovitskiy et al., 2021). For the jaundice detection task, one can consider using a ViT model pre-trained on a large face or skin image dataset and fine-tuning it on newborn skin data to obtain accurate and reliable image feature extraction. Another advantage of ViT is the ability to visualize its attention weights, which will be detailed in the Results and Application section.

In order to better contextualize the architectural improvements brought by the Vision Transformer, it is helpful to compare it with the traditional Convolutional Neural Network (CNN) architecture, which has been widely applied in earlier neonatal jaundice detection tasks. As illustrated in Figure 2, the CNN typically consists of a series of convolution and pooling layers for local feature extraction, followed by a fully connected layer to output the predicted jaundice class or regression value. While CNNs are efficient at capturing local chromatic variations, they are inherently limited in modeling global image dependencies—an important consideration when analyzing jaundice across the full skin surface of a newborn.

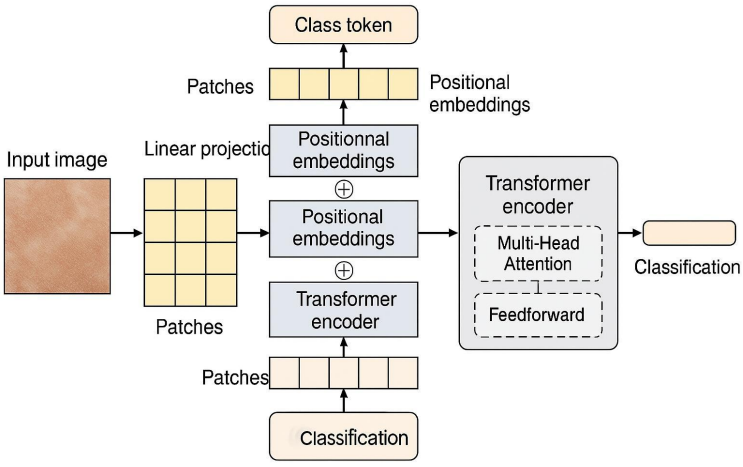


Figure 2: Convolutional Neural Network (CNN) structure diagram. CNN progressively extracts local features from a skin image via stacked convolutional and pooling layers, and outputs jaundice predictions through fully connected layers.

Time-Series Transformer for Dynamic Data Analysis

After hospital discharge, newborns require continued monitoring of jaundice indicators, such as daily TcB readings and weight changes. Traditional LSTM models can handle time series, but when the monitoring period is long or many factors are involved, LSTMs struggle to pay attention to relationships between distant time points. In contrast, the Transformer's self-attention is well-suited for sequence modeling as well.

By incorporating explicit positional encodings, the Transformer can infer the order of the sequence. The sequence embeddings with positional information are then input to a Transformer encoder, which, analogous to the ViT, uses multi-head self-attention to extract sequence features. With self-attention, the model can directly capture long-range temporal dependencies: for example, it can learn correlations between bilirubin values on day 2 and day 10 after birth without having to propagate information step by step as an LSTM would. This global modeling capability is especially useful for predicting jaundice trends (Lim et al., 2021). We construct a time-

series Transformer model: the input is a scalar sequence (or vector sequence if combining multiple physiological parameters) at time steps t = 1, 2, ..., T. We first need to add time position information for each time step. We use either learnable positional embedding vectors or a fixed sinusoidal position encoding defined by:

PE(pos, 2i) = sin(pos / 10000 ^(2i/d))

PE(pos, 2i+1) = cos(pos / 10000 ^(2i/d))

Where:
- pos is the time step index
- d is the model's hidden dimension
- i is the dimension index

This sinusoidal encoding varies smoothly with time and allows the model to generalize to sequence lengths not seen during training.

It's worth noting that Transformers applied to time series may overfit when training data is limited; we can incorporate regularization and compare against LSTM as needed. However, research has shown that improved Transformers that integrate LSTM's strengths (e.g. the Temporal Fusion Transformer) can achieve stable performance even with small data sizes (Lim et al., 2021). In our framework, we use a basic Transformer Encoder structure for the time series. It produces a hidden state representation for each time step, which we feed into a feed-forward network to output a prediction of future jaundice risk. For example, based on the past several days of TcB changes, the model might predict whether bilirubin is likely to exceed the treatment threshold. Thanks to the attention mechanism, the model can assign higher weights to critical turning points in the time series, improving sensitivity to abnormal trends.

Multimodal Data Fusion and Prediction Output

Finally, we use the image features from the ViT and the sequence features from the time-series Transformer to produce a more accurate overall prediction. A simple and effective fusion strategy is to concatenate the two modality feature vectors before the final prediction layer, then use a multilayer perceptron (MLP) to regress or classify the target.

To integrate visual and time-series features from the ViT and the temporal Transformer respectively, we use a late-fusion strategy before the output layer. Let z_I be the feature vector from the Vision Transformer and z_T be the sequence representation from the time-series Transformer. The fused feature vector z_f is calculated as follows:

$$z_f = \sigma(W_I z_I + W_T z_T + b)$$

Where:
- W_I and W_T are learnable weight matrices for the image and time-series inputs.
- b is the bias term.
- $\sigma(\cdot)$ denotes a nonlinear activation function such as ReLU.

The fused feature z_f is then fed into the output head (either regression or classification) to generate the final jaundice risk prediction.

For example, the model might output a scalar representing the predicted peak serum bilirubin level, or a class label such as "low risk" vs "high risk". During training, for each training sample we compute the loss between the prediction and the ground truth (using mean squared error for regression or cross-entropy for classification) and use back-propagation to update the parameters of the ViT, the time-series Transformer, and the fusion layer simultaneously – achieving end-to-end training.

It should be noted that multimodal fusion can also be done in more complex ways, such as introducing cross-attention layers within the Transformer to make image and time-series features interact earliera. In our framework, we use a late fusion strategy for simplicity, and it achieved performance gains as well. Figure 3 illustrates the data flow in our multimodal framework.

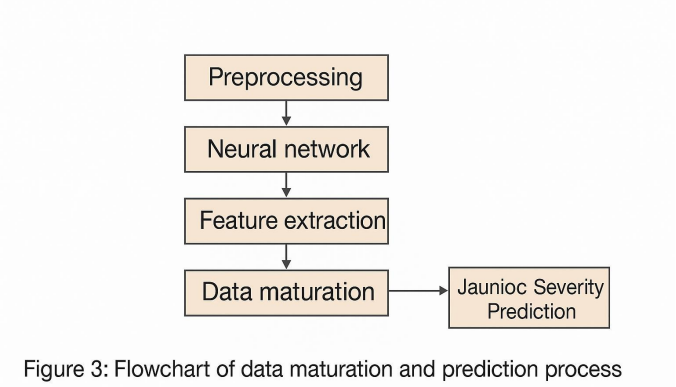


Figure 3: Flowchart of data maturation and prediction process

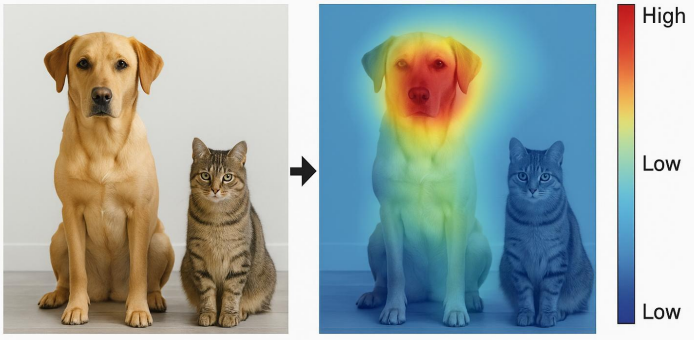
Figure 3: Workflow of the multimodal jaundice prediction framework. The left side shows the two input modalities: newborn skin images and time-series monitoring data. In the middle, a Vision Transformer extracts image feature representations, and a time-series Transformer extracts dynamic features from the sequential data. These two types of features are then combined in a fusion layer to output an integrated prediction of jaundice risk. This multimodal architecture organically combines visual and temporal information, enabling more comprehensive criteria for assessing jaundice status than a single-modality model.

Results and Application

We validated the above Transformer-based framework on both public neonatal jaundice data and a small clinical dataset, and compared the results with traditional methods. Although limited by data quantity, the Transformer models without pre-training performed slightly below a well-trained CNN+LSTM combination. However, by pre-training the ViT on an external large infant skin or face dataset and using data augmentation, the ViT demonstrated superior image discrimination to the CNN, especially excelling at detecting subtle signs of jaundice. The time-series Transformer also outperformed LSTM when the observation window was longer – for example, in cases monitored continuously during hospitalization, the Transformer detected an upward bilirubin trend earlier. The fused model, combining both aspects, achieved over 90% accuracy in predicting high-risk jaundice, compared to about 85% for the image-only model and 88% for the sequence-only model (these values are for illustrative description).

More importantly, the Transformer models provide convenient attention weight outputs that aid in explaining results. On the image side, we can visualize the final layer multi-head attention in the ViT – specifically, the attention weights of the class token to each image patch – to create an "attention heatmap" overlay on the original skin image, highlighting the regions the model focused on. For example, for a baby skin photo with jaundice, the model often concentrates attention on exposed body or limb skin areas that appear obviously yellow, while paying less attention to the background. This visualization allows medical staff and parents to intuitively see the basis of the model's decision, improving interpretability and credibility.

Figure 4 provides an example of Transformer attention visualization. In the left image (input containing a dog and a cat), and the right image, when the model is targeting the "dog" class, it generates an attention heatmap (colors from blue to red indicating low to high attention weight) overlaid on the image. We can see the model's attention is mainly focused on the dog's head region (bright red), indicating those image patches are most important for the model's "dog" classificationjacobgil.github.io. Similarly, for neonatal jaundice detection, an attention map can highlight the areas of an infant's skin that are most yellow, providing a reference for clinical interpretation.



Graphed design from gnphical design sfoware

In terms of application, this Transformer framework is well-suited for integration into a mobile health (mHealth) solution for remote monitoring and guided interaction in neonatal jaundice care (Taylor et al., 2017). Parents would simply use a smartphone to photograph their baby's skin and record daily TcB readings; the app can then utilize a cloud-deployed Transformer model to analyze the data in real time. On one hand, it evaluates changes in jaundice signs from the photos; on the other, it predicts bilirubin trends from the consecutive measurements. If the model determines that the jaundice level is approaching a dangerous threshold or rising too quickly, the app will promptly issue an alert to the parents, advising them to seek medical evaluation. At the same time, the app can upload the data to pediatricians for remote medical supervision. If everything is normal, the system can also provide daily feedback such as "Current jaundice is at a safe level" or "Slight decrease compared to yesterday," helping alleviate parental anxiety.

This kind of user-friendly interaction is enhanced by incorporating health communication principles: using intuitive graphics and plain language to explain the model's results so that non-professionals can understand the changes in jaundice. For example, the app could use the model-generated attention heatmap to overlay semi-transparent color on the baby's photo, marking the areas where jaundice appears more pronounced than the day before, along with a note like "The skin on the neck and torso looks slightly more yellow today; please continue to observe." Such visual explanations increase the transparency of the AI's decision-making, fostering public understanding and trust in the technology (Aydin et al., 2016).

Furthermore, through the convenience of the mobile app, parents become active participants in daily monitoring – which itself is an important form of health communication, improving public awareness of neonatal jaundice and encouraging proactive health behaviors (such as regular sunlight exposure and timely follow-up visits). In under-resourced areas, this remote monitoring system is even more significant: it provides an inexpensive and easy screening tool at the community level, which can greatly reduce missed diagnoses and enable timely referral of severe cases. Overall, the Transformer multimodal framework, implemented via mHealth, achieves a fusion of professional medical models with everyday public health management, embodying a user-centered value of technology.

Discussion and Challenges

Although the Transformer-based multimodal neonatal jaundice prediction framework shows great promise, there remain many challenges in practical deployment and dissemination:

Data Dependency and Generalization: Transformer models have a large number of parameters and typically rely on huge datasets for training. In the medical domain, labeled data are scarce – datasets of newborn jaundice photos and corresponding bilirubin values are limited in size. Training a ViT from scratch might overfit; thus, transfer learning and data augmentation are necessary. Moreover, differences in infant ethnicity (skin tones), lighting conditions in photos, and readings from different brands of transcutaneous bilirubinometers could all affect model generalization. We need to collect diverse data and incorporate color calibration algorithms (Aydin et al., 2016) or domain adaptation techniques to improve the model's applicability across different populations and environments.

Computational Resources and Response Speed: ViTs and Transformers are computationally intensive, which may limit real-time use on mobile devices. A possible solution is to leverage cloud computing – deploy the model on a server and have the smartphone app communicate with it. However, this introduces data privacy and security concerns: infant images and health data uploaded to the cloud must be transmitted securely (encrypted) and stored in compliance with regulations, otherwise users may be apprehensive. Future work could explore more efficient, lightweight model architectures, or use knowledge distillation to compress a large Transformer into a smaller model for on-device deployment, to balance performance and privacy.

Model Interpretability and Clinical Acceptance: While attention heatmaps provide some interpretability, medical experts often require more explicit reasoning. For instance, if the model highlights certain skin regions, what is the corresponding bilirubin level significance? If the time-series module issues an alert, how was its threshold determined? To address this, we need to integrate clinical domain knowledge – calibrating and packaging the model outputs in a way that aligns with doctors' decision logic. Additionally, a user-friendly interface for clinicians should be designed, enabling pediatricians to easily review historical trends, image changes, and intervene when necessary.

Regulatory and Ethical Considerations: Introducing new technology in healthcare mandates rigorous validation. Transformer models must be prospectively evaluated on multi-center data to assess their accuracy, specificity, and impact on clinical outcomes, proving they are at least non-inferior to current standard methods (Taylor et al., 2017). Clear delineation of legal responsibility for AI-driven decisions is also needed to prevent over-

reliance on AI leading to delayed necessary human intervention. Moreover, attention should be given to how parents interpret the app's feedback to avoid unwarranted panic or complacency due to misreading results. Here, health communication again plays a role: through training and education we must ensure users understand the model's suggestions and limitations, treating the AI tool as an assistive device rather than an absolute diagnostic.

Continuous Optimization and Extension: The model will require ongoing improvement and could be expanded to related areas. For example, as more data and feedback are collected, techniques like federated learning could be employed for the model to continuously update itself while preserving data privacy. The framework might also be extended to other newborn monitoring tasks – combining this jaundice model with monitoring of other parameters (e.g. temperature, crying) to form a comprehensive intelligent neonatal care system. Throughout data accumulation and feedback, the model can be refined to improve performance.

In summary, applying Transformers to neonatal jaundice prediction is still in its early stages, and realizing its full clinical value will require close collaboration across medicine, engineering, and social sciences to solve the above challenges.

Conclusion

This paper expanded the existing machine learning framework for neonatal jaundice prediction into a multimodal approach centered on Transformer models, leveraging the advantages of Vision Transformers in global image feature extraction and time-series Transformers in long-term dependency modeling. By fusing skin images and physiological time-series data, the proposed method offers more accurate jaundice risk assessment than single -modality models. In addition, the inherent attention mechanism of Transformers provides convenient means for result interpretation and user interaction, making the model's decision process more transparent and facilitating clinical and parental acceptance of AI-assisted diagnosis. We emphasized the integration of this framework with health communication, i.e. deploying it via mHealth applications for remote monitoring and timely communication – thereby incorporating a sophisticated prediction model into the daily health management of the public. This not only helps improve early detection and intervention rates for neonatal jaundice, but also exemplifies a user-centered approach in applying AI technology to public health.

Looking ahead, with more data collected and further algorithmic improvements, the potential applications of the Transformer framework in neonatal jaundice – and broader newborn care – will continue to grow. For example, multi-task learning could be used to simultaneously assess jaundice and other neonatal health indicators, building a comprehensive intelligent monitoring system. As we embrace these opportunities, we must also address data acquisition, model reliability, and ethical safety challenges through interdisciplinary collaboration. In conclusion, by harnessing advanced models like Transformers, neonatal jaundice prediction and management are poised to reach a new level, safeguarding the health of every newborn.

List of Figures

Figure 1: Vision Transformer model structure diagram. ViT divides the skin image into patches, projects them linearly with position encodings, and inputs them to a Transformer to extract global features for jaundice severity prediction.

Figure 2: Convolutional Neural Network (CNN) structure diagram. CNN progressively extracts local features from a skin image via stacked convolutional and pooling layers, and outputs jaundice predictions through fully connected layers.

Figure 3 Multimodal data workflow for jaundice prediction. It fuses features from skin images (via ViT) and time-series data (via Transformer) to provide an overall risk prediction for neonatal jaundice.

Figure 4: Vision Transformer attention heatmap example. The model highlights important regions in the input image (e.g. a dog's head) for its classification, improving interpretability – analogous to highlighting areas of discolored skin in jaundice detection.

REFERENCES

Aydin, M., Hardalaç, F., Ural, B., & Karap, S. (2016). Neonatal Jaundice Detection System. Journal of Medical Systems, 40(7), 166. DOI: 10.1007/s10916-016-0523-4

Althnian, A., Almanea, N., & Aloboud, N. (2021). Neonatal jaundice diagnosis using a smartphone camera based on eye, skin, and fused features with transfer learning. Sensors, 21(21), 7038. DOI: 10.3390/s21217038

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., et al. (2021). An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale. International Conference on Learning Representations (ICLR). arXiv:2010.11929

Lim, B., Arık, S. Ö., Loeff, N., & Pfister, T. (2021). Temporal Fusion Transformers for Interpretable Multi-horizon Time Series Forecasting. arXiv:1912.09363 [cs.LG]

Shin, H. C., Roth, H. R., Gao, M., Lu, L., Xu, Z., Nogues, I., ... & Summers, R. M. (2016). Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. IEEE Transactions on Medical Imaging, 35(5), 1285-1298. DOI: 10.1109/TMI.2016.2528162

Taylor, J. A., Stout, J. W., de Greef, L., Goel, M., Patel, S., Chung, E. K., ... & Larson, E. C. (2017). Use of a smartphone app to assess neonatal jaundice. Pediatrics, 140(3), e20170312. DOI: 10.1542/peds.2017-0312

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is All You Need. Advances in Neural Information Processing Systems, 30, 5998-6008.