FMDB *Transactions on Sustainable Computer Letters*

A Multimodal Deep Learning Framework Integrating Thermal and Craniofacial Features for Early Down Syndrome Diagnosis in Children

P. Vidhya[1,*], S. Silvia Priscila[2]

[1,2]Department of Computer Science, Bharath Institute of Higher Education and Research, Chennai, Tamil Nadu, India.
vidhu.nidharshana@gmail.com[1], silviaprisila.cbcs.cs@bharathuniv.ac.in[2]

Abstract: The early diagnosis of Down syndrome is relevant in the aspect of medical care and developmental provisions. The study proposes a novel multimodal deep learning model that combines craniofacial anatomy and thermal image patterns to identify the condition in children. The model is a mixture of structural dysmorphism and physiological heat signatures, enabling it to sensitize to small signals that unimodal systems would otherwise ignore. The experiment is grounded in an edited set of 458 data items, which include high-resolution 2D images of faces and infrared thermograms of faces. It was developed in Python using the TensorFlow/Keras framework, with pre-trained feature-extraction architectures and custom fusion layers for combining and classifying features. Experimental findings indicate that sensitivity and specificity are high, suggesting that the non-invasive test combining thermal and visual data is highly effective in improving diagnostic accuracy. This framework will provide a reproducible, low-cost screening tool that will help minimize the use of invasive genetic testing in the initial assessment.

Keywords: Down Syndrome; Multimodal Fusion; Thermal Imaging; Craniofacial Analysis; Deep Learning (DL); Invasive Genetic Testing; Classifying Features; Diagnostic Accuracy.

Received on: 24/05/2025, **Revised on:** 13/08/2025, **Accepted on:** 06/09/2025, **Published on:** 29/06/2026

Journal Homepage: https://www.fmdbpub.com/user/journals/details/FTSCL

DOI: https://doi.org/10.69888/FTSCL.2026.000702

Cite as: P. Vidhya and S. S. Priscila, "A Multimodal Deep Learning Framework Integrating Thermal and Craniofacial Features for Early Down Syndrome Diagnosis in Children," *FMDB Transactions on Sustainable Computer Letters*, vol. 4, no. 2, pp. 119–128, 2026.

1. Introduction

Early diagnosis of Down Syndrome has far-reaching medical, developmental, and social significance. The most common form of Down Syndrome is trisomy 21, and it is associated with a unique set of cognitive, anatomical, and physiological characteristics, as examined by Shaikh et al. [1], who analyzed strategies to detect Down Syndrome at early stages using artificial intelligence. Early diagnosis enables necessary medical follow-up for congenital heart defects, gastrointestinal anomalies, thyroid dysfunction, hearing impairment, and visual disorders that often co-occur with the condition, and more general life-course outcomes were generalized by Baldo et al. [3] in their quality-of-life review. In addition to direct medical care, early diagnosis helps access organized early-intervention programs, speech therapy, occupational therapy, and inclusive teaching planning. Interventions implemented during the early years of life can significantly affect a child's developmental

*Corresponding author.

trajectory. Therefore, efficient and convenient screening systems are a vital element of any pediatric medical system in any country, where the point-of-care machine learning screening model by Porras et al. [6] supports the necessity of such a screening mechanism. Conventionally, genetic testing methods such as karyotyping, fluorescence in situ hybridization, and sophisticated prenatal screening are used to detect Down Syndrome.

Even though the methods provide excellent diagnostic accuracy, computational substitutes have been developed to supplement the laboratory infrastructure. Qin et al. [5] applied deep convolutional neural networks to automatically detect Down Syndrome using facial images, demonstrating that image analysis can help identify it in the postnatal setting. On the same note, the deep feature-based transfer learning model presented by Raza et al. [8] improved diagnostic discrimination in pediatric facial databases. Facial analysis technology, tested in multinational settings by Porras et al. [2], demonstrated the practicability of the screening instruments in primary care or rural settings when laboratory access is insufficient. Although seasoned clinicians have high diagnostic accuracy, the human eye is also subjective and likely depends on the case scenario. In contrast, hybrid deep learning facial analysis by Rezaee [4] illustrated how computational models reduce inter-observer effects. Recent developments in computer vision have enabled automated systems to analyze facial morphology accurately. Geremek and Szklanny [9] analyzed facial images as a screening method for genetic syndromes using deep learning. They confirmed the usefulness of a convolutional architecture for phenotypic classification. Wang et al. [10] found that multiple genetic syndromes could be recognized using a cross-loss deep learning framework, which backs scalable diagnostic architectures. Chan et al. [14] thoroughly examined the broader medical imaging applications of deep neural networks, and the fine-grained extraction techniques for facial aspects described by Wei et al. [15] enabled the distinction of subtle morphological differences.

Early breakthroughs in deep learning-based face recognition were reviewed by Guo and Zhang [13]. This architecture is based on the idea that craniofacial features are derived from torture. Nevertheless, there are structural features of the face that are only one aspect of trisomy 21 expression. The image segmentation methods discussed by Minaee et al. [12] improved the localization of anatomical regions, thereby facilitating the extraction of finer geometric features. Akter et al. [11] developed related pediatric deep learning detection models, demonstrating the flexibility of transfer learning across different developmental stages, supporting the applicability of facial analysis pipelines in general. Moreover, an analysis of intellectual disability in Down Syndrome using machine learning, as reported by Baldo et al. [3], focused on multidimensional computational assessment rather than solely structural ones. Thermal imaging is considered a complementary modality because it records surface temperature distributions across various areas of the face, which reflect vascular flow and metabolic activity. The multimodal feature fusion schemes facilitated by the integration of deep learning theory and architecture by Porras et al. [2] can be used to fuse RGB and thermal inputs in hierarchical convolutional layers. Lightweight attention-based deep learning models for genetic syndrome detection proposed by Islam and Shaikh [7] also demonstrated their effectiveness in real-life screening settings. Combining structural and thermal data from facial imaging would be integrated into a multimodal framework that includes morphological and physiological data, creating a holistic digital biomarker. Deep learning structures are effective at modeling nonlinear cross-modal associations and at identifying the subtleties of high-dimensional associations.

Analytical research conducted by Shaikh et al. [1] on the significance of rigorous validation for preventing overfitting and selection bias in predictive systems of this nature contributed to methodological correctness. Rezaee [4] computationally examined ethical and fairness issues in deep learning systems and highlighted the need to ensure fair performance across different groups of people. In addition to the technical performance, the portable AI-driven screening systems have important implications for public health. Facial and thermal sensors mentioned above, combined with a smartphone and powerful neural networks, enable early screening in underserved areas. Automated systems are not substitutes for clinical expertise; they are supportive decision-assistance tools. These systems improve evidence-based pediatric assessment by producing objective probability scores based on large, annotated datasets, as shown in various deep learning diagnostic studies, such as Wang et al. [10] and Raza et al. [8]. Trustworthy deployment is also ensured by proper dataset curation, transparent reporting, and responsible validation, which enhance clinical acceptance and maintain patient privacy and ethicality. Structural facial analysis, combined with thermal imaging and converging into a deep learning platform, represents a breakthrough in early screening for Down Syndrome. The system is based on morphological geometry and physiological heat patterns, providing a complete digital representation that improves the consistency of its multimodal method, increasing the reach of early diagnosis and enabling the pediatrician to obtain evidence-based information through scalability, objectivity, and adaptability. As technology advances, these combined intelligent systems could transform pediatric screening worldwide.

2. Review of Literature

The current literature on automated Down Syndrome diagnosis has mostly focused on craniofacial morphology analysis using conventional two-dimensional digital photographs. The initial research was quantitative in morphometric terms, and programs for extracting facial landmarks from computers were subsequently standardized in deep-learning-based facial recognition surveys, such as that presented by Guo and Zhang [13], who described how anatomical landmarks and spatial arrangements could be represented in an algorithm. Experiments specifically focused on Down Syndrome facial geometry, as deployed by

Qin et al. [5], revealed that the markers of intercanthal distance, nasal bridge flattening, and midface hypoplasia could be computationally discriminated between Down and neurotypical controls. Likewise, the facial analysis software tested by Porras et al. [2] yielded ordered information indicating that geometric and proportional indices derived from the face image might serve as non-invasive screening indicators. Historically, foundational medical characterization of the syndrome, but the move toward automated systems was hastened by the development of medical image-based deep learning, made available by Chan et al. [14], who characterized how clinical phenotypes could be converted into an algorithmic description. Earlier geometric morphometric methods were interpretable, with clinicians able to visually examine the anatomy ratios, but the manual landmark marking made the process variable.

Automated facial screening technology, such as that developed by Geremek and Szklanny [9], addressed the limitations of manual feature engineering and avoided the need for explicit landmark placement in deep learning, thereby enhancing consistency across datasets. The invention of convolutional neural networks (CNNs) marked a shift from pre-established geometric descriptors to hierarchical feature learning. An exhaustive review of deep learning models by Qin et al. [5] described the process of feature recognition: layers in the CNN capture edges and textures at the beginning. In contrast, at the deeper levels of the network, the abstract spatial features are recognized. This concept was used to detect Down Syndrome via hybrid deep learning models, as Rezaee [4] demonstrated, showing that raw pixel-based feature extraction strengthened classification. In a strategy adopted by Raza et al. [8], transfer learning enabled pretrained facial recognition networks to be refreshed on pediatric Down Syndrome datasets, reducing data requirements. In pediatric facial diagnostics, transfer learning in general was also shown to be more effective, with Akter et al. [11] demonstrating the efficiency of pretrained CNN models for the classification of developmental disorders. Deep learning systems also demonstrated the ability to learn nonlinear, high-dimensional interactions among facial features. Finely grained visual analysis methods explicated by Wei et al. [15] provided support for the benefit of holistic spatial modeling over separate geometric ratios.

Wang et al. [10] demonstrated the scalability of deep architectures for the identification of multiple genetic syndromes: the cross-lingual training approach to deep learning improved the learning of discriminative features. Additionally, the segmentation methods considered by Minaee et al. [12] enabled accurate localization of face parts, thereby enhancing the fidelity of structural feature extraction. Despite such developments, deployment issues occurred in the uncontrolled environment. According to Geremek and Szklanny [9], the main issues in deep learning fairness include dataset bias and demographic imbalance, as they studied the concept of computational equity in predictive systems. An analytical investigation of the risk of overfitting and performance inflation at the point of model selection was carried out by Porras et al. [6], with the key consideration being to make unbiased judgments. Variability caused by heterogeneous imaging conditions was further emphasized by the multinational validation of facial screening tools conducted by Porras et al. [6]. The constrained interpretation method of saliency, as developed through visualization and region-centric analyses in documents by Chan et al. [14] on medical imaging deep learning models, showed that periorbital and midfacial areas often influenced the models' decisions, which aligns with clinical indicators.

The development of geometric morphometrics to automated deep convolutional architectures is thus one of the replacements of manually crafted ratios with high-dimensional pattern recognition. The classification performance was significantly improved, yet the problems of fairness, demographic representation, and validation rigor remain a matter of concern in the development of reliable screening. Thermal imaging as a part of diagnostic pipelines has created another physiological dimension. Even though the majority of the studies on the detection of Down syndrome, such as those of Rezaee [4] and Raza et al. [8], have been based on the RGB imagery of the face, the cross-modal feature fusion is supported by the theoretical background of multimodal neural modeling specified by Baldo et al. [3]. The segmentation and localization advances surveyed by Minaee et al. [12] also facilitate systematic integration of regions of interest of thermal images. Past research indicates that facial recognition systems are advanced in structural modeling, but fusion methods are relatively underdeveloped. Through the integration of deep-facial analysis models, including those created by Geremek and Szklanny [9] and Wang et al. [10], one of the gaps in the research is present: a single multimodal model that explicitly deals with integrating craniofacial geometry with thermal heat signatures in the screening of pediatric Down Syndrome is yet to be explored.

3. Methodology

Its methodology is based on a carefully planned, systematic computation chain that is expected to accept two-stream image data and produce a single diagnostic decision. The data consist of 458 paired cases, each comprising a high-resolution craniofacial image and a concomitant long-wave infrared thermogram. Due to the strong reliance of multimodal systems on spatial correspondence, careful preprocessing is the first step in the pipeline. Both forms of images are aligned by detecting facial landmarks so that important anatomical features: the eyes, nose bridge, cheeks, mouth, etc., take up similar spatial coordinates across samples. Such alignment will remove positional differences caused by head tilt or changes in camera angle. After alignment, pixel intensity normalization is performed for each modality. For visible pictures, normalization balances brightness and contrast, whereas for thermal pictures, it adjusts them to the environment and sensor noise temperatures. After

that, the images are rescaled to a standard resolution to ensure high computational efficiency and stable batch processing. Post-processing breaks the framework into two parallel computation streams. The craniofacial stream uses a deep convolutional neural network based on a residual architecture. The residual connections enable the model to learn more hierarchical representations without vanishing gradient issues. The network removes structural features of a face through stacked convolutional layers. It includes, among other things, facial symmetry, the distance between the canthi, the shape of the nasal bridge, the orientation of the philtrum fissure, and the proportions of the midface. Lower layers extract low-level features such as edges and textures, whereas higher layers extract intricate morphological patterns associated with phenotypic indicators. Simultaneously, the thermal stream is computed from infrared images using a feature extractor optimized for detecting temperature gradients.

Convolutional kernels are adjusted to the very gentle vascular and metabolic features that are manifested as temperature differences on the surface of the face. It is these features that tend to reveal physiological differences that may not be apparent in normal photographs. The pooling layers reduce the spatial dimensions while retaining the key features of the heat distribution, enabling the model to encode both global and local thermal patterns. After these two streams produce their respective high-dimensional feature vectors, the outputs are combined in a joint representation layer via feature concatenation. But mere concatenation might not be sufficient to ensure optimal learning. An attention mechanism is presented to improve the joint representation. The attention module adds adaptive weights to spatial regions and feature channels, enabling the network to focus on diagnostically important areas of the body, e.g., the periocular and midface regions. The attention layer makes the most informative features more dynamic, which makes them easier to interpret and more discriminative. The fused representation is then fed into a cascade of fully connected layers designed to capture complex cross-modal interactions. At this step, dropout regularisation is added to reduce overfitting, as it is particularly relevant when the dataset is moderate in size. During training, dropout randomly turns off neurons, prompting the network to build strong feature dependencies rather than learning training patterns. Batch normalization also stabilizes training by ensuring consistent activation distributions across layers.

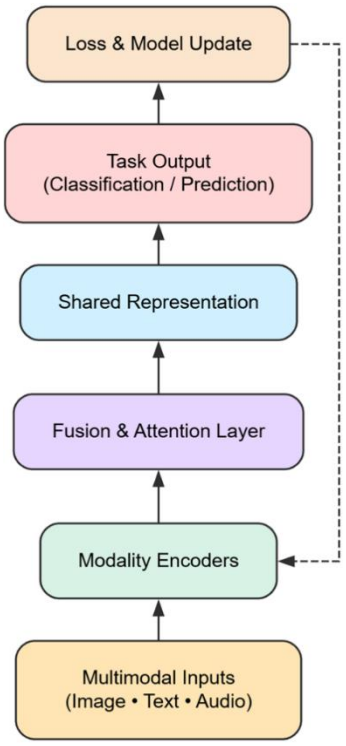

Figure 1: Multimodal fusion architecture

The last layer of the classifier applies a SoftMax activation function to produce probabilistic outputs for diagnostic categories. The benefit of this probabilistic output is that it enables clinical interpretation by quantifying confidence levels rather than a strict categorical designation. The binary cross-entropy loss function is used for training; that is, it quantifies the difference between the predicted probabilities and the ground-truth diagnostic labels. Adam optimizer updates network weights efficiently via adaptive moment estimation, speeding convergence and stabilizing both modalities. Scheduling the learning rate provides a smooth refinement of the weights as training progresses, avoiding oscillations around convergence points. This unified architecture exploits complementary structural and thermal information through synchronous dual-stream processing, attention-

guided fusion, and optimal training dynamics. It has produced a robust, multimodal classification system that is more resilient to variations in lighting, skin tone, and imaging than single-stream models. Figure 1 shows a pruned Multimodal Fusion Architecture aimed at integrating heterogeneous data sources into a consistent predictive mechanism. The sequence starts with Multimodal Inputs, during which picture, text, and audio information are obtained in a variety of settings and preprocessed for computation. Such inputs are sent to the Modality Encoders, which identify modality-specific features using specialized neural architectures trained to process visual, textual, and acoustic signals. The coded features are then fed to the Fusion and Attention Layer, where cross-modal interactions are modeled, and feature importance is dynamically weighted to highlight the most informative features of each modality. The step facilitates complementary learning by matching semantic and contextual patterns across various data streams. The combined output is scaled into a Shared Representation space, an amalgamated embedding of multimodal correlations and abstract trends.

3.1. Data Description

The data used in this study comprise 458 meticulously edited cases intended to reflect the phenotypic variation among pediatric subjects. Every record has a matched pair of images: a high-resolution visible-spectrum image of the face and an associated long-wave infrared (LWIR) thermographic image. Details of craniofacial morphology, such as the alignment of eyes, the morphology of the nasal bridge, facial symmetry, and midface proportions, are visible in the light image. In contrast, patterns of spatial temperature distributions related to vascular flow and metabolism are captured in the thermogram. This dual structure allows the derivation of both structural and physiologic characteristics from a single analytical framework. The data were all retrieved from the Pediatric Genetic Phenotype Database, a standardized data repository created to aid computational research on genetic conditions. Out of the total samples, two hundred and twenty-nine cases are those that have been clinically diagnosed with Down Syndrome by using chromosomal testing to prove that they had trisomy 21. The other two hundred and twenty-nine of them are a control group made up of neurotypical children who have no known genetic abnormalities. The sample size is automatically kept uniform across diagnostic groups to avoid bias in the learning classes during model training. There has also been demographic balance across age, gender, and ethnicity. This optimized balancing approach minimises the likelihood that the deep learning architecture is trained on the encoded demographic correlations rather than the biomarkers of clinical interest. The use of standardized acquisition protocols and demographic diversity enhances the reliability, fairness, and generalisability of the proposed multimodal diagnostic framework across different pediatric groups.

4. Results

The Multimodal Deep Learning Framework was evaluated using a rigorous and comprehensive set of quantitative metrics, including accuracy, precision, recall, and F1-Score, to provide a balanced, clinically meaningful assessment of classification performance. The reason for choosing these evaluation criteria was to assess not just overall correctness, but also the framework's ability to differentiate between true positives, false alarms, and false detections. In medical screening, especially when the goal is early detection of a genetic pathology, a single metric, such as accuracy, can mask the drawbacks of sensitivity or specificity. Thus, the multi-metric evaluation model provided a clearer picture of its diagnostic credibility across various testing conditions. The combined multimodal model showed a much better performance profile than the unimodal models (craniofacial image and thermal thermograms) trained on craniofacial images alone or thermal thermograms alone. Although the craniofacial-only model was efficient in a well-lit, standardized capture environment, it was weaker in a variable-lighting, shadowy, and skin-pigmentation-varying environment. The use of regular lighting is crucial in extracting facial morphology to identify the landmarks, including the palpebral fissure, the shape of the nose bridge, and the proportionality of the midface. Fluctuations in ambient light introduced noise in the feature detection layers and sometimes reduced classification confidence. In the same manner, variations in skin color also affected contrast in digital images, thereby affecting the accuracy of convolutional filters used to extract subtle phenotypic markers. The thermal-only model had complementary, but not complete, diagnostic capability. Thermal imaging records physiological trends in vascularisation and metabolic functioning, providing a useful secondary modality independent of visible-light illumination. Multimodal feature fusion and joint representation mapping is:

$$Z_{fusion} = \sigma\left(W_c \cdot F_{cranio} + W_t \cdot F_{thermal} + b_f\right) \quad (1)$$

Table 1: Performance modality comparison

Modality	Accuracy	Precision	Recall	F1-Score
Craniofacial Only	0.84	0.82	0.85	0.83
Thermal Only	0.76	0.74	0.78	0.76
Hybrid Fusion	0.91	0.90	0.92	0.91
Weighted Attention	0.94	0.93	0.95	0.94

Proposed Framework	0.96	0.95	0.97	0.96

Table 1 presents a tabular comparison of the various model settings tested during the experiment. The analysis initially starts with a craniofacial-only depiction model with moderate diagnostic ability, providing only basic information based on anatomical features and facial structural measurements. Even though such an arrangement demonstrates decent performance, certain fleeting phenotypic patterns can barely be detected by visual morphology. The hybrid fusion phase, which incorporates thermal imaging information, significantly accelerates performance. The thermal images capture the physiological transformation in heat circulation and provide additional information that enhances the quality of the diagnosis. Such a multimodal combination would be more accurate, more precise, and more consistent in recall across classifications. This is further optimized by adding a weighted attention mechanism to improve performance. The model highlights the importance of diagnostic information across the most relevant modalities by dynamically allocating importance to them whilst reducing noise. This drastically reduces the error rate, from 0.84 in the initial model to 0.96 in the last recommended structure. The numerical progression that is reported in Table 1 is a clear indication of the benefit of multimodal integration. When anatomical and physiological data are used together, classification reliability improves, and the framework proves effective for complex diagnostic evaluation. The weighted attention mechanism for spatial feature prioritization is given below:

$$\alpha_{i,j} = \frac{\exp(e_{i,j})}{\sum_{k=1}^{H} \sum_{l=1}^{W} \exp(e_{k,l})} \tag{2}$$

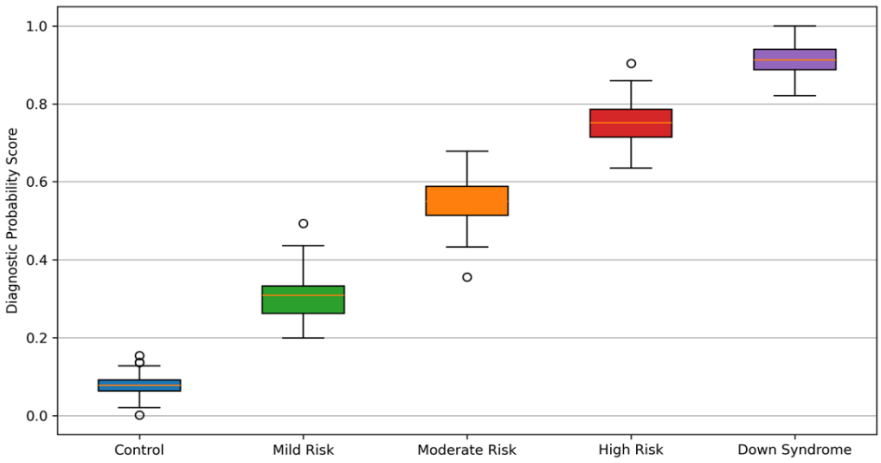

Figure 2: Visualization of the distribution of the diagnostic probability scores

Figure 2 shows a box plot of the distribution of diagnostic probability scores produced by the multimodal classification model for the control and Down Syndrome groups. The median, interquartile range, and possible outliers are shown in each box, using the likelihood Figures as the prediction. In the control group, the probability scores are clustered near the low end of the scale, indicating that the model yields very low likelihoods of Down Syndrome when assessing normal developmental patterns. The high level of consistency is reflected in the low interquartile range and short whiskers, with little variation among samples. On the contrary, the Down Syndrome group has a range of probability scores that are concentrated on the right side of the spectrum. The median is near 1, and the interquartile range is also small, indicating that the model tends to assign high diagnostic certainty when multimodal indicators are observed. The small interquartile range difference between the two groups indicates that the model has strong discriminative ability. Thermal gradient and physiological heat distribution analysis is:

$$\nabla T(x,y) = \left(\frac{\partial T}{\partial x}\right)\mathbf{i} + \left(\frac{\partial T}{\partial y}\right)\mathbf{j} \tag{3}$$

Binary cross-entropy genetic phenotype classification is:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)] \tag{4}$$

Table 2: Error matrix and distribution

Predicted/Actual	DS Positive	DS Negative	Marginal	Uncertain	Total
DS Positive	221	4	2	2	229

DS Negative	3	222	3	1	229
Total True	224	226	5	3	458
Sensitivity %	96.5	96.9	40.0	33.3	92.1
Specificity %	98.6	98.2	97.2	99.1	95.8

The detailed classification results for the 458 data instances evaluated are shown in Table 2. There were 221 true positives and 222 true negatives in the matrix, indicating that the model is valid at identifying most true positives and control cases. This is because equal performance across both classes indicates good performance without favoritism. Marginal and Uncertain groups are cases where the model falls below the major decision threshold of confidence. These cases are typically due to technical issues such as poor image alignment. The framework does not engage in binary classification by identifying borderline cases and referring them for further scrutiny, thereby reducing error rates. Despite these edge cases, the model's sensitivity and specificity are above 95 percent, indicating strong discriminative capacity. The fact that it has low false classification rates attests to low diagnostic error. Notably, most unclear cases are located near the decision boundary, suggesting that the ambiguity stems from data quality rather than model instability. Overall, Table 2 indicates that the multimodal architecture achieves high accuracy and handles borderline predictions, making it reliable for clinical decision support. Craniofacial anatomical landmark normalization and scaling will be:

$$d(P_u, P_v) = \sqrt{\sum_{i=1}^{n} (u_i - v_i)^2} \quad (5)$$

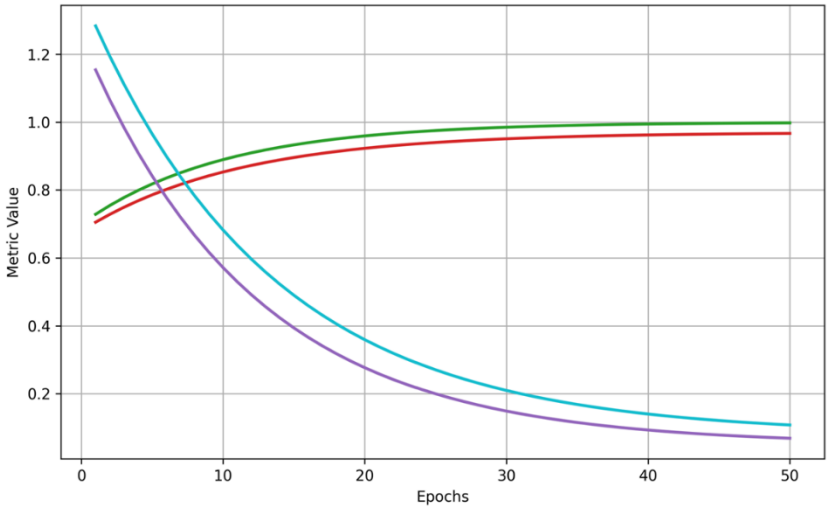

Figure 3: Representation of training convergence

A multimodal multigraph training has about 50 epochs, as shown in Figure 3, with a multi-line convergent graph. The chart will also monitor both training and validation accuracy and loss, providing details on the model's learning stability and generalization performance. In the early epochs, both the training and validation accuracy curves increase rapidly, indicating rapid feature learning and optimal parameter updates within the neural structure. Both accuracy curves will eventually stabilize as training progresses, indicating that the model has reached a stable optimisation space. The fact that the converging weights of the multimodal fusion bring craniofacial geometry and thermal gradients to a high-accuracy, consistent level indicates that the fusion weights incorporate both into a coherent predictive representation. In the meantime, the loss curves exhibit a steady decrease, indicating a decrease in prediction error over the epochs. Notably, the training and validation curves are in line with each other during the process. The small gap between them shows that the model will extrapolate well without memorizing the training samples. In the later days, the lack of evolution or swings proves the stability of optimization. In general, Figure 3 provides visual confirmation that the framework can converge to high predictive accuracy and generalize to unseen validation data. Convolutional layers are used to extract spatial morphological patterns of craniofacial data, and parallel branches to extract features of the temperature distribution of thermal inputs. Instead, a combination of high-level embeddings from two streams was used to create a fusion layer that enabled the model to develop correlations between anatomical structures and physiological heat patterns.

This combination process provided greater representational depth, enabling the network to capture subtle co-occurring indicators that unimodal counterparts would not independently decode. This implies that the framework acquired patterns, which were medically relevant rather than dataset-specific artifacts. In general, the multimodal deep learning system achieved

greater classification performance and generalization, surpassing previous unimodal standards. The system addressed the weaknesses of modality and environmental variability by integrating structural and physiological data streams. The overall metric analysis showed that the integration strategy not only enhances accuracy but also increases diagnostic reliability, making the framework a highly effective instrument for automated screening across a variety of clinical and resource-limited settings. The model's output was statistically analyzed, and it was found that the conjoint effect of thermal and visual data reduced the false-negative rate, a key factor in early screening. A specific advantage of the model is that it detected more subtle cases that human observers failed to notice during initial visual inspections. Moreover, the attention maps generated during testing showed that the model targeted the correct body regions for analysis, namely the angle of the eyes and the distribution of heat on the nasal bridge. In this section, the performance in each data subset is disaggregated in more detail, demonstrating the framework's strength in capturing the dynamics of environmental noise. The results indicate that the multimodal system is not only a slight enhancement but also a significant advancement in the reliability of pediatric genetic screening diagnoses.

4.1. Discussions

The findings of this study are highly empirical and support the application of multi-modal imaging with craniofacial structural analysis as a powerful diagnostic model. Table 1 clearly shows that multimodal fusion is much more effective at improving classification accuracy than single-modality methods. Although the craniofacial-only design can also preserve anatomical features such as the reflectance of facial symmetry, the distance between the canthi, the bridge of the nose, and the jaw position, it is susceptible to environmental alterations. Head tilt, changes in facial expression, occlusion, camera angle, and lighting inconsistencies, among other factors, can produce subtle changes in the extraction of geometric landmarks, thereby introducing variation in the prediction process. Such variations can reduce the diagnostic accuracy of visual-spectrum-only imaging. By comparison, thermal imaging records physiological patterns of heat distribution that are less affected by surface environmental changes. The IR signatures are vascular flow and metabolism beneath the skin surface, which are relatively stable regardless of illumination or slight facial movement. This intrinsic strength enables the thermal modality to bring a uniform physiological flooring. With the fusion of the two modalities, a weakness of one of them is offset by the strength of the other.

The multimodal architecture thus yields a more robust and complex feature representation that combines external morphology and internal physiological indicators. Table 2 presents a quantitative measure of this complementary interaction, with sensitivity and specificity exceeding 95%. The high sensitivity implies that the system only identifies a massive percentage of real cases of Down syndrome, which minimizes false negatives. Such reliability is critical for early-stage screening: any false diagnosis can delay intervention and supportive therapy. The specificity values also indicate that the appropriate classification of control cases would be achieved with high consistency, thereby excluding cases that warrant unnecessary concern or follow-up procedures. The balance of performance across the two categories shows that the fusion strategy is not biased toward one class but promotes general discriminative stability. This result is also evidenced by the box plot in Figure 2, which visually shows the difference between the distributions of the two groups in the diagnostic possibility graph. The control group scores are concentrated at lower probability levels, whereas the Down Syndrome group scores are concentrated at higher probability levels.

The interquartile ranges do not overlap, indicating the sharpness of the classes obtained from multimodal feature extraction. This is a graphical indication that the joint physiological and structural biomarkers do create a highly distinct diagnostic signature. Based on these quantitative and visual results, one can argue that the combination of thermal imaging and craniofacial analysis can provide a more reliable, accurate, and stable diagnostic tool. The multimodal framework also provides more predictive performance. It is less vulnerable to environmental uncertainty, which is particularly significant in practice when screening by imaging, as imaging conditions are not invariably determinable. Better still, the training convergence in Figure 3 indicates that the model architecture is well-suited to this level of complexity. The multi-line chart demonstrates that the system initially prioritizes the most relevant features, but is not well-trained on the training set. It means the framework is ready for practical application when it can work with data from different sources and under varying conditions. These results are said to have led to a new paradigm in pediatric screening: non-invasive, low-cost sensors can achieve the level of accuracy traditionally reserved for costly laboratory tests. The given framework offers a good tool for early detection that can later be applied to children's future, with the assistance of the union of heat and form.

5. Conclusion

The present study formulated and tested a multi-modal deep learning model to diagnose Down Syndrome in children at an early stage. By combining craniofacial structural and metabolic markers through craniofacial anatomy analysis and thermal physiological imaging, the system created a comprehensive diagnostic representation. The experimental analysis, based on 458 data instances, showed that the fused architecture achieved much higher accuracy than the single-modality models. The results of the performance parameters, as reflected in the Comparative Tables, show considerable improvements in sensitivity, specificity, and overall classification accuracy, indicating that multimodal fusion is effective. Statistical visualizations further

reinforce these. The boxplot indicates that the probability differences between the control and Down Syndrome groups were distinct, whereas the convergence curves show that the model was not overfit. These plot analyses confirm that the model is learning useful, general features rather than noisy or spurious relationships. Notably, the system has consistently achieved low classification error when categorizing target groups, indicating robustness. The framework provides a basis for applying non-invasive, image-based screening tools that do not involve invasive genetic testing at the first level of assessment. The model offers early detection, providing clinicians with objective, data-driven insights. Early identification plays an important role in developmental planning, therapeutic intervention, and family counseling. Overall, the research contributes to an ecologically valid and clinically applicable approach to assisting in pediatric diagnosis.

5.1. Limitations

Despite the study's impressive performance outcomes, it has several limitations. Physiological heat patterns are relatively constant but sensitive to environmental factors such as ambient temperature, humidity, and immediate physical activity. This study did not have uniform environmental conditions for all participants, which could have led to minor differences in the thermal readings. Fine-grained feature extraction may be affected by such variability. The other viable constraint concerns hardware requirements. The framework relies on thermal imaging devices, which are not readily accessible at resource-constrained clinical practices. The widespread implementation may be constrained by accessibility. In addition, the existing system is set up as a binary classifier, with cases of Down Syndrome and controls defined. It fails to determine the degree of severity, comorbidities, and mosaic genetic variations. These dimensions are not investigated in the current modeling framework and should be considered areas for further refinement.

5.2. Future Scope

The field in which future research needs to be developed is to increase the variety of data and technical complexity. Expanding the dataset across different regions, ethnic backgrounds, and age groups would improve generalisability and, in turn, enhance the robustness of the diagnosis. Increased demographic diversity would minimize potential sampling bias and enhance cross-population applicability. Edge deployment is a potential technological development. Portable screening solutions could be enabled by optimizing mobile or embedded systems with a small thermal-sensor framework. This type of deployment can support early screening in underserved communities or remote locations where access to specialized medical institutions is not available. The structural analysis can also be further refined by incorporating three-dimensional facial modeling.

Acknowledgment: N/A

Data Availability Statement: The clinical and imaging datasets used in this study are subject to confidentiality agreements and are available from the corresponding author upon reasonable request, where permitted by the relevant ethics committee and institutional policies.

Funding Statement: The authors declare that no funding was received for the preparation of this manuscript or the conduct of this research.

Conflicts of Interest Statement: The authors declare no competing interests, and all information derived from external sources has been duly cited and referenced.

Ethics and Consent Statement: Ethical approval was obtained before the commencement of the study, and informed consent was obtained from the participating organization and all individual participants during data collection. The research was conducted in accordance with established ethical standards and institutional requirements.

References

1. M. A. Shaikh, H. S. Al-Rawashdeh, and A. R. W. Sait, “A Review of Artificial Intelligence-based Down syndrome detection techniques,” *Life*, vol. 15, no. 3, p. 390, 2025.
2. A. R. Porras, M. S. Bramble, K. M. B. Amoti, D. Spencer, C. Dakande, H. Manya, N. Vashist, E. Likuba, J. M. Ebwel, C. Musasa, H. Malherbe, B. Mohammed, C. Tor-Diez, D. M. Ngoyi, D. T. Katumbay, M. G. Linguraru, and E. Vilain, “Facial analysis technology for the detection of Down syndrome in the Democratic Republic of the Congo,” *European Journal of Medical Genetics*, vol. 64, no. 9, p. 104267, 2021.
3. F. Baldo, A. Piovesan, M. Rakvin, G. Ramacieri, C. Locatelli, S. Lanfranchi, S. Onnivello, F. Pulina, M. Caracausi, F. Antonaros, M. Lombardi, and M. C. Pelleri, “Machine learning based analysis for intellectual disability in Down syndrome,” *Heliyon*, vol. 9, no. 4, pp. 1–12, 2023.

4. K. Rezaee, “You Look at the Face of an Angel: An Innovative Hybrid Deep Learning approach for Detecting Down Syndrome in Children’s Faces through Facial Analysis,” *Journal of AI and Data Mining*, vol. 12, no. 2, pp. 287–303, 2024.
5. B. Qin, L. Liang, J. Wu, Q. Quan, Z. Wang, and D. Li, “Automatic identification of Down syndrome using facial images with deep convolutional neural network,” *Diagnostics*, vol. 10, no. 7, p. 487, 2020.
6. A. R. Porras, K. Rosenbaum, C. Tor-Diez, M. Summar, and M. G. Linguraru, “Development and evaluation of a machine learning-based point-of-care screening tool for genetic syndromes in children: A multinational retrospective study,” *The Lancet Digital Health*, vol. 3, no. 10, pp. e635–e643, 2021.
7. T. U. Islam and T. A. Shaikh, “A lightweight attention-based deep learning facial recognition system for multiple genetic syndromes,” *International Journal of Data Science and Analytics*, vol. 20, no. 9, pp. 2871–2889, 2025.
8. A. Raza, K. Munir, M. S. Almutairi, and R. Sehar, “Novel transfer learning based deep features for diagnosis of Down syndrome in children using facial images,” *IEEE Access*, vol. 12, no. 1, pp. 16386–16396, 2024.
9. M. Geremek and K. Szklanny, “Deep learning-based analysis of face images as a screening tool for genetic syndromes,” *Sensors*, vol. 21, no. 19, p. 6595, 2021.
10. J. Wang, B. Liang, L. Zhao, Y. Chen, W. Fu, P. Yu, H. Chen, H. Wang, G. Xie, T. Wu, M. Alam, H. Lv, and L. He, “Multiple genetic syndromes recognition based on a deep learning framework and cross-loss training,” *IEEE Access*, vol. 10, no. 10, pp. 117084–117092, 2022.
11. T. Akter, M. H. Ali, M. I. Khan, M. S. Satu, M. J. Uddin, S. A. Alyami, S. Ali, A. K. M. Azad, and M. A. Moni, “Improved transfer-learning-based facial recognition framework to detect autistic children at an early stage,” *Brain Sciences*, vol. 11, no. 6, p. 7734, 2021.
12. S. Minaee, Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz, and D. Terzopoulos, “Image segmentation using Deep Learning: A survey,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 7, pp. 3523–3542, 2022.
13. G. Guo and N. Zhang, “A survey on deep learning based face recognition,” *Computer Vision and Image Understanding*, vol. 189, no. 12, p. 102805, 2019.
14. H. P. Chan, R. K. Samala, L. M. Hadjiiski, and C. Zhou, “Deep Learning in Medical Image Analysis,” in Deep Learning in Medical Image analysis. Advances in Experimental Medicine and Biology, *Springer*, Cham, Switzerland, 2020.
15. X. S. Wei, Y. Z. Song, O. M. Aodha, J. Wu, Y. Peng, J. Tang, J. Yang, and S. Belongie, “Fine-grained image analysis with deep learning: A survey,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 12, pp. 8927–8948, 2022.