

Determining the Health Condition of Newborns Using Apgar Score and Thermal Camera Imaging

Farhad Nassehi^{1,2}, Doga Kaya¹, Pelin Kocak¹, Emine Bilun Haskaraman¹, Seray Akbag³, Osman Eroglu¹

¹Biomedical Engineering, TOBB University of Economics and Technology, Ankara, Turkey

²OKUMAN Medikal Sistemler A.Ş., Ankara, Turkey

³Bilkent City Hospital, Ankara, Turkey

(f.nassehi, d.kaya, p.kocak, e.haskaraman, erogul)@etu.edu.tr, serayakbag@gmail.com

Abstract—APGAR (Appearance, Pulse, Grimace, Activity, Respiration) Score is a system that evaluates the post-birth health status of newborns and is scored between 0-10. This score, which consists of five parameters: appearance, pulse, grimace, activity and respiration, is obtained by scoring each parameter between 0-2. The condition of newborns with an APGAR score of 7-10 is considered reassuring, a score of 4-6 is considered moderately abnormal and a score of 0-3 is considered low and critical. This study aims to automatically determine the APGAR score of newborn babies with thermal camera imaging and artificial intelligence support and to classify the newborn according to the APGAR score. For this purpose, data was collected from a total of 22 newborns whose gestational age was between 28-42 weeks. Thermal images of newborns were preprocessed in the MATLAB software environment and first the babies were separated from the background. After this stage, important features determined as a result of the literature research were extracted. The extracted features and the APGAR scores obtained from nurses, which are accepted as gold standards, were used in the Python software environment for classification. Stacking Classifier models were used in the regression and classification stages. Among the classification algorithms, the most successful model in distinguishing risky and healthy babies was the Stacking Classifier with 88.59% (± 0.0264) accuracy. The proposed method may assess APGAR score non-invasively, non-contact, objectively and early detection of risky newborns was provided.

Keywords—Neonatal; Thermal imaging; Machine learning; Apgar score, Death risk

I. INTRODUCTION

Newborns' behaviors and movements are critical to understand. Because these movements and behaviors are the language of the newborn and provide important information about the baby. A careful post-birth examination helps to detect anomalies, cardiovascular disorders or birth injuries that may pose a risk to the newborn's successful adaptation to life outside the uterus. Every newborn should have a short physical examination within the first few minutes after birth [1]. Basic assessment methods, such as the APGAR (Appearance, Pulse, Grimace, Activity, Respiration) score, help to detect abnormalities in the baby's vital functions [2]. The APGAR score is a quick method used by healthcare providers to assess a

newborn's condition immediately after birth. The APGAR score takes into account 5 components: Appearance, Pulse, Grimace, Activity, Respiration. Each parameter receives a value of 0, 1 or 2 and thus the APGAR score is assessed out of 10 points. In all newborns, scoring is done in the first minute and fifth minutes after birth. Babies with an APGAR score of less than 7 at fifth minutes are also evaluated at tenth minutes.

TABLE 1. APGAR SCORE PARAMETERS

	0	1	2
Appearance	The body is blue or pale in color	Pink body with blue or pale extremities	Normal, pink color all over the body
Pulse beats/min	No pulse	<100	>100
Grimace	No response to stimulation	Some facial grimace in response to stimulation	Coughing, sneezing, or response to stimulation
Activity	No movement	Only some movement of the extremities	Active and spontaneous movement

A score between 7 and 10 is considered reassuring, a score between 4 and 6 is considered moderately abnormal and a score between 0 and 3 is considered low in infants [3,4]. The APGAR score is considered a useful predictor of neonatal death. Evidence from studies suggests that 50% of preterm (≤ 31 weeks) infants with an APGAR score ≤ 3 die in the neonatal period (within the first 28 days after birth), compared with less than 2% of preterm infants with an APGAR score ≥ 7 . In term infants (≥ 37 weeks), 10% of infants with APGAR score ≤ 3 and 0.2% of term infants with APGAR score ≥ 7 died. When all these data were analyzed, it was observed that 42.2% of neonatal deaths occurred in infants with an APGAR score ≤ 3 and almost 20% of all newborns with an APGAR score ≤ 3 died in the neonatal period [5,6].

At present, since the APGAR score method is based on observation, it may show variability among healthcare personnel. Infant nurses with different levels of expertise may give different APGAR scores to the same infant. This situation causes both loss of time and inconsistencies in evaluation. The findings obtained from the studies showed that the level of

knowledge about APGAR score evaluation among nurses and midwives is insufficient. A study found that more than 90% of nurses continued APGAR scoring even in cases with missing data, and that APGAR scores obtained by different nurses were low to medium level compatible and therefore had low reliability [7]. A study conducted in Turkey found that midwives did not know all of the APGAR score evaluation criteria. In particular, they have limited knowledge about the evaluation of heart rate and muscle tone [8]. The use of probes and sensors that require contact when determining the APGAR score causes discomfort, hygiene risk and stress on the sensitive skin of infants. The non-contact, non-radiating and non-invasive nature of thermal imaging devices offers a safe and comfortable use for newborns [9,10].

This study proposes a novel framework to determine newborns' health status immediately after birth, in a non-invasive and non-contact way using thermal camera imaging and artificial intelligence methods. In this context, it aims to develop a machine learning model to prevent the APGAR score from varying according to nurses' assessments and to reduce the workload of healthcare personnel. By using thermal images of newborn babies and machine learning algorithms, it is aimed to provide an objective, fast and non-contact assessment of the baby's health status.

II. MATERIALS AND METHODS

A. Participants

Newborns delivered via Caesarean section in the Neonatal Intensive Care Unit (NICU) of Ankara Bilkent City Hospital were included in this study, with no gender-based exclusion criteria. A total of 305 thermal images were collected from 22 newborns (15 male, 7 female) of varying gestational ages to ensure a balanced dataset. The study was approved by the Ethics Committee of Ankara Bilkent City Hospital. APGAR scores were categorized as poor (≤ 6) or good (≥ 7).

B. Data Acquisition

Thermal imaging was performed using a FLIR® C2 7400 high-performance thermal camera ($\pm 0.1^\circ\text{C}$ sensitivity, 320×240 resolution), capturing detailed surface temperature distributions. Following maternal informed consent, standardized environmental conditions (temperature, lighting, airflow) were maintained during data acquisition. Thermal images were obtained at the 1st, 5th, and when necessary, 10th minute after birth, focusing on regions such as the abdomen, back, and face, with care taken to avoid obstructions. Concurrently, nurses recorded clinical APGAR scores, serving as gold-standard labels.

C. Image Processing

The image preprocessing pipeline involved segmenting the newborn from the background and isolating relevant temperature data. Initially, raw thermal images were cropped to focus on the infant while excluding irrelevant temperature points. The thermal camera's temperature scale was indexed to $34\text{--}38^\circ\text{C}$ ($\pm 2^\circ\text{C}$ from the standard body temperature of 36°C), minimizing background influence. Cropped images were converted to grayscale and enhanced using histogram equalization to improve contrast and highlight subtle temperature variations. The resulting images exhibited clearer

edge definition, facilitating the separation of the newborn from the background. Black pixels (with a value of zero) indicating the background were detected and used to create a mask, which was applied to the original cropped image. This process effectively isolated the newborn's temperature profile, ensuring that only temperature data corresponding to the infant were retained for subsequent feature extraction.

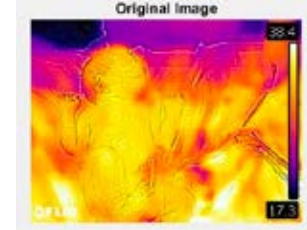


Fig 1. Original Thermal Image of the Newborn

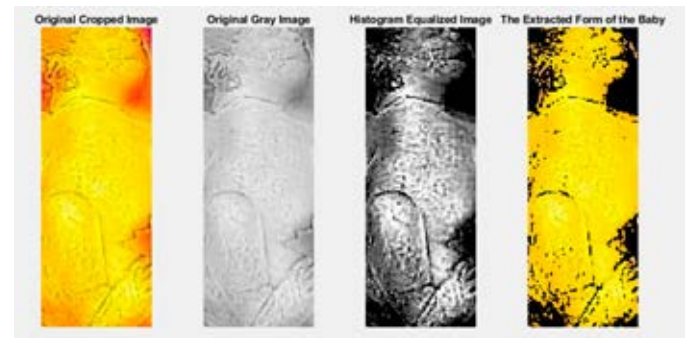


Fig 2. Results of Background Elimination

D. Feature Extraction

In the feature extraction stage, numerical features representing the thermal characteristics of the newborns were derived from the preprocessed images to predict health statuses of newborns. Both time-domain and frequency-domain (FD) features were utilized. Time-domain (TD) features included mean, standard deviation, variance, median, skewness, interquartile range (IQR), and root mean square (RMS). For the frequency domain, a two-dimensional Fast Fourier Transform (2D-FFT) was applied to the images, from which statistical features such as mean, standard deviation, variance, minimum, maximum, median, skewness, IQR, RMS, MAD, energy, and kurtosis were extracted. Features were also computed from the RGB color channels and pixel intensity values of the thermal images. In total, 40 features were generated, encapsulating both spatial and frequency information of the temperature distribution, thereby enhancing the model's ability to distinguish between different APGAR score categories.

E. Classification Algorithms

In this study, newborns were classified into two groups based on their APGAR scores: critical (≤ 6) and healthy (≥ 7), comprising 6 and 16 newborns, respectively. The dataset was split into 70% training and 30% testing using the hold-out method. To optimize model performance, the training set was further processed using GridSearchCV for hyperparameter tuning (5-fold cross-validation), LASSO for feature selection, and SMOTE for addressing class imbalance. GridSearchCV

systematically searched for optimal hyperparameters across five classification models, while LASSO ($\lambda=0.01$) reduced the feature set from 40 to 12 key features. SMOTE ($k=5$) was applied to synthetically augment the minority class (critical APGAR). Decision tree (DT), random forest (RF), support vector machine (SVM), multi-layer perceptron (MLP), and stacking classifier algorithms were used as classifiers in this study. The final models were trained and evaluated across 50 iterations, with performance reported as mean and standard deviation for metrics such as accuracy, sensitivity, specificity, and Receiver Operating Characteristic (ROC), and Area Under Curve (AUC).

III. RESULTS

A total of 40 features were extracted, and 12 of them were selected using the LASSO method. The selected FD features are mean, variance, and MAD. The selected TD features are mean, variance, median, skewness, median (red channel), RMS (red channel), mean (blue channel), median (blue channel), and RMS (blue channel).

In the result of GridSearchCV; for DT, (*criterion* = entropy, *max_depth* = 10), for RF, (*n_estimator* = 200, *max_depth* = 20), for SVM (*C* = 10, *kernel* = rbf), for MLP (*hidden_layers_size* = [100,50], *max_iter* = 1000) were selected as hyperparameters of classifiers. Stacking Classifier is a model that is a combination of three models (RF, SVM, MLP), the previously selected hyperparameters of these three models were used in the hyperparameter selection. Therefore, the hyperparameters of the logistic regression (LR) model used in the last part of the stacking classifier were selected as (*C* = 10, *solver* = liblinear).

The Stacking Classifier model shows the best performance among all models. This model has performed better than the other models with 88.59% (± 0.0264) accuracy, 85.04% (± 0.0411) sensitivity and 92.14% (± 0.0344) specificity values. Since Stacking Classifier is a model created by combining different models, this situation shows that ensemble methods increase the classification success. MLP, on the other hand, has become the model that shows the best performance after Stacking Classifier, especially with 93.42% (± 0.0357) specificity value. When all models are examined, it is seen that the model that shows the lowest success in terms of all metrics is Decision Tree. When the ROC Curve and AUC values are examined, the most successful model in separating the risky and healthy babies with an AUC value of 0.95 is Stacking Classifier. Random Forest comes after Stacking Classifier with an AUC value of 0.94, while MLP and SVM values are successful models with AUC values of 0.92. Decision Tree again showed the lowest success with an AUC value of 0.78. As a result, when all metrics are evaluated, it is seen that the model that can most successfully classify newborns in critical condition requiring medical intervention and newborns in healthy condition is Stacking Classifier.

TABLE 3. CLASSIFICATION RESULTS

Classifiers	Accuracy	Sensitivity	Specificity
RF	85.38% (± 0.0317)	80.99% (± 0.0549)	89.78% (± 0.0465)
DT	76.53% (± 0.0389)	72.27% (± 0.0679)	80.79% (± 0.0572)
SVM	84.63% (± 0.0274)	76.30% (± 0.0486)	92.96% (± 0.0349)
MLP	87.45% (± 0.0293)	81.48% (± 0.0526)	93.42% (± 0.0357)
Stacking Classifier	88.59% (± 0.0264)	85.04% (± 0.0411)	92.14% (± 0.0344)

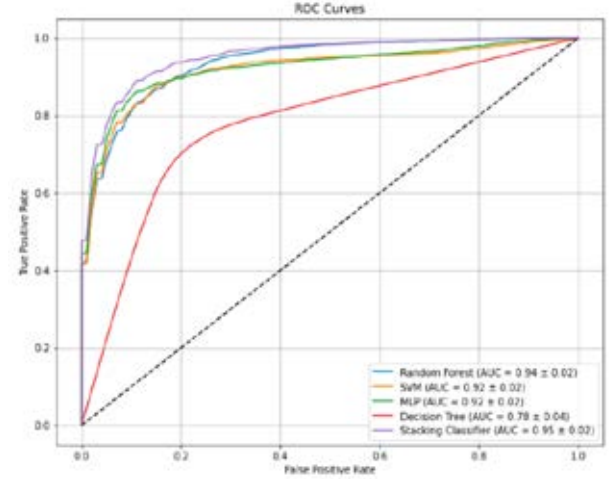


Fig 3. ROC Curve and AUC Values of Classifiers

IV. DISCUSSION

According to Table 3. and Fig. 3 Stacking classifier perform better than other classifier algorithms. Stacking classifier reached average 88.59% accuracy, 85.05% sensitivity, and 92.14% specificity rates. Ervural and Ceylan (2022) focuses on the classification of neonatal disease by analyzing the limited thermal image data. The study is meant to obtain high-accuracy classification with small and unbalanced datasets using convolutional neural networks (CNNs) and data augmentation methods. 20 thermal images were collected from each of 44 newborns, for a total dataset of 880 thermal images. The highest accuracy (94.55%) was obtained by training the model with 15 of the 20 images of each infant. In addition, when the model was tested without training, a classification accuracy of 89% was observed, which is more meaningful in terms of the real-life performance of the model [10]. Jung and his colleagues focused on determining muscle activation during resistance exercises with the help of infrared thermal imaging. Observing the three major muscles in the upper arm (biceps brachii, triceps brachii, and deltoid muscle) different weights, exercise durations, and muscle activation sequences were used in the experiments. Thermal images were recorded with a FLIR thermal camera and heat maps were generated through processing the images. In order to classify muscle activation a CNN model was trained with the obtained dataset. The CNN model was able to classify muscle activations with 92.3% accuracy. Results show that thermal imaging has the potential to assess muscle isolation and exercise accuracy [11].

According to our knowledge no study was found on the direct determination of APGAR score in newborns with a thermal camera. However, a patented study is noteworthy in this context among the studies conducted on the determination of APGAR score in the literature. The study aims to automatically determine the APGAR score using the body temperature, muscle tone, pulse, voice and respiratory physiological data of newborns. Body temperature sensor, respiratory sensor, pulse sensor, microphone and pulse oximeter are used to obtain data from newborns.

V. CONCLUSION

This thesis aims to determine the APGAR score, which is used to evaluate the health status of newborns after birth, using thermal cameras and machine learning algorithms. In addition, it aims to identify babies with critical health conditions and ensure immediate medical interventions. The biggest challenge encountered during the project was that the number of newborns with low APGAR scores remained in the minority. Since the majority of the data collected in the hospital consisted of healthy and high APGAR score babies, a sufficient amount of data on low-scoring newborns could not be obtained, and an unbalanced data set was formed. An important technological development is provided in the estimation of APGAR score and the use of thermal cameras in the field of health. The ability to determine the APGAR score with the body temperature values of the newborn with the help of a thermal camera allows faster and more objective data to be provided to healthcare professionals. This situation standardizes the assessment of newborn health, especially in the first hours after birth, and accelerates medical intervention decisions of healthcare personnel. In addition, this system can increase the survival chances of risky babies by allowing early detection of babies with critical health conditions. Additionally, the ease of use of the system and the ability to directly estimate the APGAR score provide advantages in delivery rooms by reducing the workload of healthcare personnel. There are certain limitations that should be considered. Firstly, the biggest limitation of this study is the small data set. This reduces the generalizability of the study. In addition, the small number of newborns with low APGAR scores also caused class imbalance. This imbalance attempted to be reduced by generating artificial data using the SMOTE method, but this still did not correspond to the real-world data. Another limitation is that the data were collected from a single hospital. Creating the dataset using the evaluations of only a few nurses from a single hospital limits the generalizability of the project. This study is a preliminary study for the continuation and sustainability of the project. The dataset is being expanded by continuing data collection from different hospitals.

VI. FUTURE WORK

As a feature work, we decided to validate our algorithm with real abnormal data and then use this algorithm as an embedded form in OKM 730 - Radiant Warmer that is produced by OKUMAN MEDİKAL SİSTEMLER A.Ş.



Fig 4. Radiant Warmer

ACKNOWLEDGMENT

This study supported by TÜBİTAK 2209-B University Students Industrial Research Projects Support Program (Application No: 1139B412402440) and OKUMAN Medikal Sistemler A.Ş.

REFERENCES

- [1] Eryürek, D., Başdaş, Ö., Korkmaz, Z., Yıldız, İ., Mumcu, Ö., & Bayat, M. (2021). Prematüre bebek annelerinin yenidoğan yoğun bakım deneyimleri. Hacettepe Üniversitesi Hemşirelik Fakültesi Dergisi, 8(2), 196-202.
- [2] Genel, T. S. B. H. S., Çocuk, M., & Başkanlığı, E. S. D. (2019). Temel Yenidoğan Bakımı. Erişim tarihi, 25.
- [3] Simon, L. V., Hashmi, M. F., & Bragg, B. N. (2017). APGAR score.
- [4] Michel, A. (2022). Review of the Reliability and Validity of the Apgar Score. *Advances in Neonatal Care*, 22(1), 28-34.
- [5] Li F, Wu T, Lei X, Zhang H, Mao M, et al. (2013) The Apgar Score and Infant Mortality. *PLOS ONE* 8(7): e69072. <https://doi.org/10.1371/journal.pone.0069072>
- [6] Cnattingius, S., Norman, M., Granath, F., Petersson, G., Stephansson, O., & Frisell, T. (2017). Apgar score components at 5 minutes: risks and prediction of neonatal mortality. *Paediatric and perinatal epidemiology*, 31(4), 328-337.
- [7] Michel, A. D., & Lowe, N. K. (2023). Accuracy and interrater agreement of registered nurses' assignment of Apgar scores to standardized clinical vignettes. *Clinical Nursing Research*, 32(3), 452-462.
- [8] Tezel, A., Özkan, H., & Çelebioğlu, A. (2005). EBELER "APGAR DEĞERLENDİRMESİ" Nİ NE KADAR BİLİYORLAR?. *Sağlık Bilimleri Dergisi*, 14(3), 163-171.
- [9] Topalidou, A., Ali, N., Sekulic, S., & Downe, S. (2019). Thermal imaging applications in neonatal care: a scoping review. *BMC pregnancy and childbirth*, 19, 1-14.
- [10] Ervural, S., & Ceylan, M. (2022). Classification of neonatal diseases with limited thermal Image data. *Multimedia Tools and Applications*, 81(7), 9247-9275.
- [11] Jarlier, S., Grandjean, D., Delplanque, S., N'diaye, K., Cayeux, I., Velazco, M. I., ... & Scherer, K. R. (2011). Thermal analysis of facial muscles contractions. *IEEE transactions on affective computing*, 2(1), 2-9.