Bigint Computing Journal
Volume 4, Issue 2, 2026, pp. 74-78
DOI: https://doi.org/10.55537/bigint.v4i2.1856
ISSN: 3032-5374
https://journal.aira.or.id/bigint/index

OCR-LSTM-Based Detection of Pork-Derived Non-Halal Ingredients from Food Labels

Muhammad Siddik Hasibuan, Suhardi, Bagus Ageng Alfahri*

Department of Computer Science, Faculty of Science and Technology, Universitas Islam Negeri Sumatera Utara, Medan, Indonesia

ABSTRACT
Packaged food labels may contain technical and multilingual ingredient terms that complicate preliminary screening for pork-derived non-halal substances. This study develops a web-based pipeline that integrates optical character recognition (OCR), automatic translation, and Long Short-Term Memory (LSTM) text classification. A public dataset of 528,092 labeled ingredient records, comprising 291,920 halal and 236,172 pork-related non-halal records, was used for model development. After text normalization, tokenization, and sequence padding, the data were divided into training, validation, and testing subsets using an 80:10:10 ratio. The final test set contained 52,810 records. The confusion matrix contained 29,182 true negatives, 12 false positives, 41 false negatives, and 23,575 true positives, corresponding to 99.90% accuracy, 99.95% precision, 99.83% recall, and a 99.89% F1-score. The web implementation accepts label images, extracts text, translates non-English content, and applies the trained classifier. The reported metrics evaluate the text classifier rather than the complete OCR-to-classification pipeline; therefore, the system should be treated as a preliminary screening tool and not as a substitute for formal halal certification.
KEYWORDS: Halal Ingredient Screening, OCR, LSTM, Food-Label Text Classification, Deep Learning

*Corresponding author: Bagus Ageng Alfahri, bagusageng073@gmail.com

Article History: Received: 2026-08-01; Accepted: 2026-08-15

1. INTRODUCTION

Halal authentication is an important component of consumer protection because pork and porcine derivatives can enter processed foods through meat, fat, gelatin, flavoring, enzymes, or other compound ingredients. Analytical studies therefore emphasize the need for reliable detection of pig-derived materials and for transparent ingredient information [1], [2], [3], [4]. For consumers, however, laboratory authentication is not available at the point of purchase, and multilingual or highly technical ingredient labels can make preliminary screening difficult.

Automated text analysis can complement, but not replace, laboratory and certification procedures. OCR converts label images into machine-readable text, while modern scene-text research shows that image quality, layout, illumination, and typography remain important sources of recognition error [5]. Once text is extracted, neural text-classification methods can model contextual relationships among ingredient terms [6], [7], [8]. LSTM networks are particularly designed to retain information across sequential inputs and remain a well-established baseline for sequence modeling [9], [10].

Previous work has begun to connect artificial intelligence with halal-food analysis. HALALCheck combines multiple intelligent components for packaged-food recognition [11], while a recent BiLSTM approach classifies halal status from ingredient lists [12]. These studies demonstrate the feasibility of AI-assisted screening, but a practical gap remains between image acquisition and ingredient-level classification when users must still provide clean text or when multilingual labels require additional processing.

This study addresses that gap by integrating OCR, automatic translation, and an LSTM classifier in a web-based workflow for pork-related ingredient screening. The contribution is an end-to-end prototype that accepts a food-label image, converts the visible ingredient list to text, normalizes non-English content, and applies a binary classifier trained on a large labeled ingredient dataset. The study also distinguishes classifier-level accuracy from end-to-end system validity, because OCR and translation errors were not independently quantified. The intended use is preliminary decision support rather than formal halal certification.

2. METHODOLOGY

2.1. Dataset and Experimental Design

The model-development dataset was obtained from the public List of Food Ingredients with Halal Label repository [13]. It contains 528,092 labeled records, including 291,920 halal records and 236,172 pork-related non-halal records. The published labels were adopted as provided and were not independently re-annotated in this study. The records were divided into 80% training data, 10% validation data, and 10% testing data. The test partition therefore contained 52,810 records and was used only after model fitting.

This design evaluates the ability of the classifier to discriminate the two labels represented in the source dataset. It does not establish coverage of every possible non-halal substance, and it does not by itself validate the OCR or translation stages. These distinctions are important because halal authentication research includes broader chemical, molecular, and species-level verification problems that cannot be reduced to ingredient-text classification [1], [2], [3], [4], [14], [15].

2.2. Text Preprocessing and LSTM Classification

Ingredient strings were standardized through text cleaning, tokenization, label encoding, and sequence padding. The tokenizer mapped ingredient terms to integer indices and the sequences were padded to a common length. The classifier used a learned embedding representation followed by stacked LSTM layers and dropout regularization. The source implementation specifies 50-dimensional token embeddings, LSTM layers with 100 and 50 units, a dropout rate of 0.2, and a dense classification head. This architecture follows the general use of recurrent neural networks for sequential text representations [6], [9], [10].

Training used the Adam optimizer [16] with a learning rate of 0.0001 and a batch size of 32. A maximum of 100 epochs was allowed, while early stopping with a patience of five epochs and model checkpointing were used to retain the best validation model. The selected run converged after 23 epochs. Accuracy, precision, recall, and F1-score were calculated from the held-out test labels, with the pork-related non-halal class treated as the positive class. These metrics are standard for binary text classification, but they should be interpreted together because high overall accuracy can conceal asymmetric false-positive and false-negative behavior [7], [8].

2.3. OCR, Translation, and Web Integration

The trained text classifier was integrated into a Flask-based web application. Users upload an image containing a product ingredient label, after which OCR extracts visible text. Non-English text can be translated into English through the Google Cloud Translation service [17]. The extracted text is then passed through the same tokenization and padding steps used during model development before the classifier returns a binary screening result. This architecture separates image acquisition, language normalization, and sequence classification so that each stage can be independently improved in future work.

3. RESULTS AND DISCUSSION

3.1. Training Behavior and Test Classification

The training curve shows rapid convergence during the early epochs. Training accuracy approaches 100%, while validation accuracy stabilizes near 99.9%. Figure 1 presents the accuracy trajectory over the 23 completed epochs.

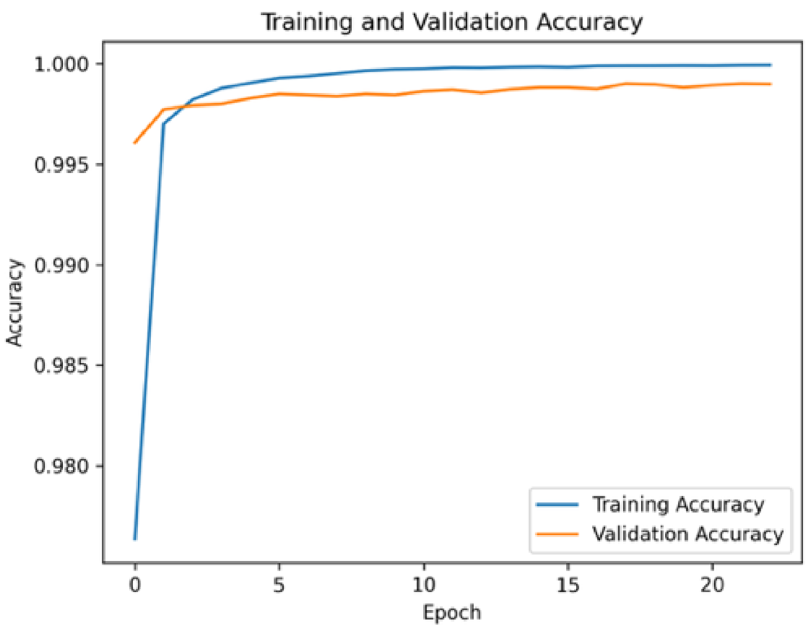

Figure 1. Training and Validation Accuracy Across 23 Epochs

The loss trajectory provides a more cautious view of generalization. Training loss continues to decline, whereas validation loss reaches a low level and then shows a modest increase during later epochs. Figure 2 therefore suggests mild late-epoch overfitting despite the very high validation accuracy, which supports the use of early stopping and best-checkpoint selection.

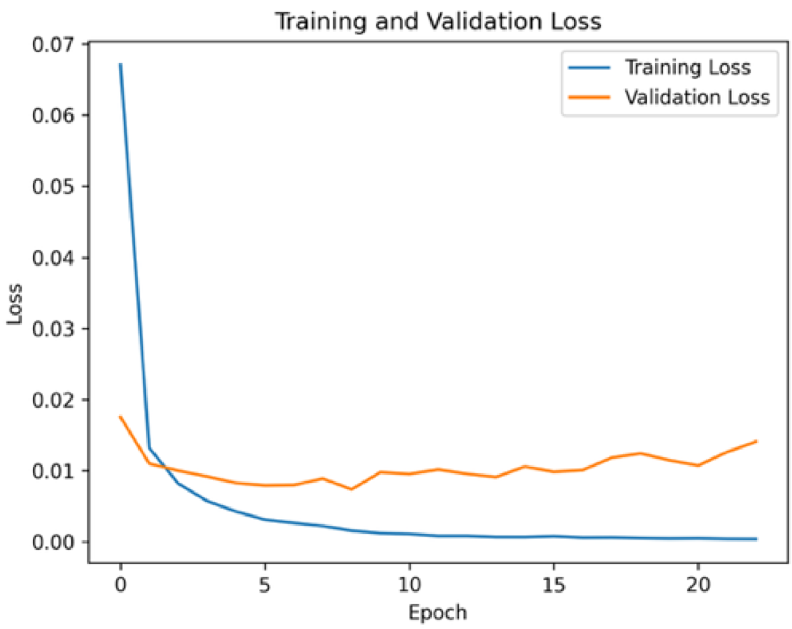

Figure 2. Training and Validation Loss Across 23 Epochs

On the 52,810-record test set, the confusion matrix contained 29,182 true negatives, 12 false positives, 41 false negatives, and 23,575 true positives. These counts yield the classification performance summarized in Table 1. The small number of false negatives is particularly relevant because a false-negative outcome would label a pork-related ingredient record as halal within the binary dataset.

Table 1. Classification Performance on the Held-Out Test Set

Metric	Value
Accuracy	99.90%
Precision	99.95%
Recall	99.83%
F1-score	99.89%

3.2. Web-Based Screening Interface

The web prototype integrates image upload, OCR-based text extraction, automatic translation, and LSTM classification in a single user workflow. Figure 3 shows the initial interface, where users can upload a food-label image and initiate analysis. The screen is localized in Indonesian for the intended user context, while the analytical pipeline operates on the extracted ingredient text.

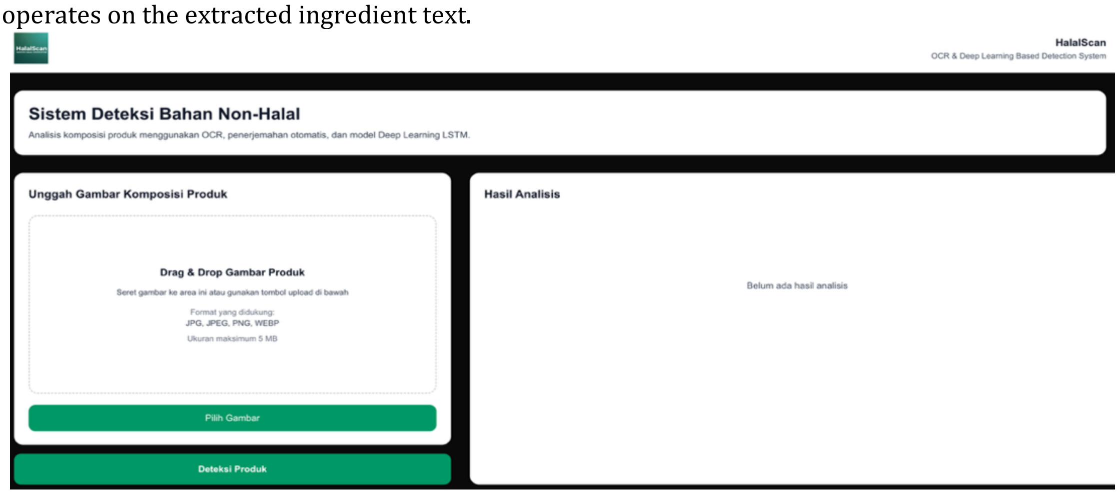

Figure 3. Web Interface for OCR-LSTM Ingredient Screening

The implementation demonstrates technical integration rather than full system-level validation. The reported 99.90% accuracy is obtained from the pre-labeled text test set, not from a separate benchmark of photographed product labels. Consequently, an OCR error, an incorrect translation, or an ingredient absent from the training vocabulary can still propagate to the final screening result even when the underlying text classifier performs well.

3.3. Comparative Discussion and Methodological Limitations

The classifier-level results are consistent with the strong sequence-modeling capacity reported for LSTM architectures [9], [10] and with broader evidence that deep neural methods can perform very well on text classification when large labeled datasets are available [7], [8]. The system-level contribution differs from ingredient-list-only approaches by connecting image input to downstream classification, which is closer to the workflow required by end users. At the same time, OCR research shows that recognition quality can deteriorate under blur, perspective distortion, low contrast, complex layouts, and small typography [5].

Several limitations constrain interpretation of the near-perfect classification metrics. The labels originate from one public dataset and were not independently audited; random record-level splitting may also place highly similar ingredient strings in both training and test partitions, which can make generalization appear stronger than it would be on unseen commercial products. The current label space covers pork-related non-halal ingredients rather than the full range of halal-compliance issues discussed in authentication research [1], [2], [3], [4], [14], [15], [18]. In addition, OCR accuracy, translation fidelity, and end-to-end product-level performance were not separately measured.

Future validation should therefore use a product-level external dataset with photographs captured under realistic conditions, deduplicate or group related ingredient strings before data splitting, and report stage-specific error rates for OCR, translation, and classification. Broader coverage should include additional non-halal or doubtful ingredients and should be evaluated against authoritative halal standards. Such extensions would move the prototype from text-classification proof of concept toward a more robust consumer-facing screening system.

4. CONCLUSION

This study developed a web-based pipeline that integrates OCR, automatic translation, and LSTM classification for preliminary screening of pork-derived non-halal ingredients on food labels. The classifier was trained using 528,092 labeled ingredient records and evaluated on a 52,810-record held-out test set. The resulting confusion matrix contained 29,182 true negatives, 12 false positives, 41 false negatives, and 23,575 true positives, corresponding to 99.90% accuracy, 99.95% precision, 99.83% recall, and a 99.89% F1-score.

The practical contribution is the integration of image-based input and text classification within a single web interface. Nevertheless, the reported metrics describe the text classifier rather than complete OCR-to-decision accuracy, and the source labels were not independently revalidated. The system is also restricted to pork-related ingredients and should not be interpreted as a substitute for formal halal certification. External product-level validation, grouped or deduplicated data splitting, broader ingredient categories, and quantitative evaluation of OCR and translation are required before higher-stakes use.

ACKNOWLEDGMENT

The authors thank Universitas Islam Negeri Sumatera Utara for the academic support provided during this research.

DECLARATIONS

Artificial Intelligence Use

Generative AI was used only to assist with language editing during manuscript revision. The authors reviewed the final text and remain responsible for the article content, analysis, and conclusions.

Author Contributions

Muhammad Siddik Hasibuan, Suhardi, and Bagus Ageng Alfahri contributed to the conception, methodology, analysis, manuscript preparation, review, and approval of the final manuscript.

Funding

This research received no external funding.

Conflict of Interest

The authors declare no conflict of interest.

Ethics Approval / Ethical Clearance

Not applicable. This study used a public labeled ingredient dataset and did not involve human participants or animals.

Data Availability Statement

The dataset used in this study is publicly available in the Kaggle repository cited in Reference [13]. Additional implementation materials may be obtained from the corresponding author upon reasonable request.

REFERENCES

[1] K. Nakyinsige, Y. B. Che Man, and A. Q. Sazili, "Halal authenticity issues in meat and meat products," Meat Science, vol. 91, no. 3, pp. 207–214, 2012, doi: 10.1016/j.meatsci.2012.02.015

[2] A. Rohman and Y. B. Che Man, "Analysis of pig derivatives for halal authentication studies," Food Reviews International, vol. 28, no. 1, pp. 97–112, 2012, doi: 10.1080/87559129.2011.595862

[3] I. Usman et al., “Advances and challenges in conventional and modern techniques for halal food authentication: A review,” Food Science & Nutrition, vol. 12, no. 3, pp. 1430–1443, 2024, doi: 10.1002/fsn3.3870
[4] A. M. Fathima, L. Rahmawati, A. Windarsih, and Suratno, “Advanced halal authentication methods and technology for addressing non-compliance concerns in halal meat and meat products supply chain: A review,” Food Science of Animal Resources, vol. 44, no. 6, pp. 1195–1212, 2024, doi: 10.5851/kosfa.2024.e75
[5] S. Long, X. He, and C. Yao, “Scene text detection and recognition: The deep learning era,” International Journal of Computer Vision, vol. 129, pp. 161–184, 2021, doi: 10.1007/s11263-020-01369-0
[6] Y. Goldberg, “A primer on neural network models for natural language processing,” Journal of Artificial Intelligence Research, vol. 57, pp. 345–420, 2016, doi: 10.1613/jair.4992
[7] K. Kowsari, K. J. Meimandi, M. Heidarysafa, S. Mendu, L. Barnes, and D. Brown, “Text classification algorithms: A survey,” Information, vol. 10, no. 4, art. no. 150, 2019, doi: 10.3390/info10040150
[8] S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, and J. Gao, “Deep learning-based text classification: A comprehensive review,” ACM Computing Surveys, vol. 54, no. 3, art. no. 62, pp. 1–40, 2021, doi: 10.1145/3439726
[9] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997, doi: 10.1162/neco.1997.9.8.1735
[10] K. Greff, R. K. Srivastava, J. Koutnik, B. R. Steunebrink, and J. Schmidhuber, “LSTM: A search space odyssey,” IEEE Transactions on Neural Networks and Learning Systems, vol. 28, no. 10, pp. 2222–2232, 2017, doi: 10.1109/TNNLS.2016.2582924
[11] S. Tarannum, M. S. Jalal, and M. N. Huda, “HALALCheck: A multi-faceted approach for intelligent halal packaged food recognition and analysis,” IEEE Access, vol. 12, pp. 28462–28474, 2024, doi: 10.1109/ACCESS.2024.3367983
[12] D. Mulyana, D. H. F. Pratama, D. Fujianti, D. A. Nurhakim, and D. Rabbani, “Automation of halal food classification using Bidirectional Long Short-Term Memory on ingredients list,” Khazanah Journal of Religion and Technology, vol. 3, no. 2, pp. 36–44, 2025, doi: 10.15575/kjrt.v3i2.1774
[13] I. A. Habibi, “List of Food Ingredients with Halal Label,” Kaggle dataset, 2023. https://www.kaggle.com/datasets/irfanakbarihabibi/list-of-food-ingredients-with-halal-label
[14] R. M. H. Raja Nhari, J. H. Soh, N. F. Khairil Mokhtar, N. A. Mohammad, and A. Mohd Hashim, “Halal authentication using lateral flow devices for detection of pork adulteration in meat products: A review,” Food Additives & Contaminants: Part A, vol. 40, no. 8, pp. 971–980, 2023, doi: 10.1080/19440049.2023.2242955
[15] A. A. Aida, Y. B. Che Man, and R. Son, “Detection of pig derivatives in food products for halal authentication by polymerase chain reaction–restriction fragment length polymorphism,” Journal of the Science of Food and Agriculture, vol. 87, no. 4, pp. 569–572, 2007, doi: 10.1002/jsfa.2699
[16] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in Proc. International Conference on Learning Representations, 2015, doi: 10.48550/arXiv.1412.6980
[17] Google Cloud, “Cloud Translation documentation,” 2026. [Online]. Available: https://cloud.google.com/translate/docs
[18] S. Siska, M. I. Jumadil, S. Abdullah, D. Ramadon, and A. Mun’im, “ATR-FTIR and chemometric method for the detection of pig-based derivatives in food products: A review,” International Food Research Journal, vol. 30, no. 2, pp. 281–289, 2023, doi: 10.47836/ifrj.30.2.01