Secure and Explainable AI for Diabetic Foot Complication Detection and Angiosome-Level Localization from Plantar Thermograms

S. Vijayaragavan[1*], S. Sivaguranathan[2]
[1,2]*Computer Science and Applications, The Gandhigram Rural Institute, Dindigul*
vijaysarainn@gmail.com, svgrnth@gmail.com

Abstract

Diabetic foot ulcers are preceded by changes in plantar skin temperature that infrared thermography can capture without contact. Machine learning models can screen these thermograms, but three issues limit their clinical use: their decisions are opaque, small input perturbations can change their outputs, and models trained on patient data can leak information about those patients. This paper proposes a secure and explainable framework for detecting diabetic foot thermal abnormality and localizing it to plantar angiosomes. From the public Plantar Thermogram Database (167 subjects: 122 with diabetes, 45 controls), 55 numerical features are extracted from the whole foot and four angiosomes of each foot. A multilayer perceptron is trained with projected gradient descent (PGD) adversarial training and differentially private stochastic gradient descent (DP-SGD), and SHAP values explain each prediction and are aggregated per angiosome for localization. Over 5 × 5-fold cross-validation, the proposed model reached an accuracy of 0.758, F1-score of 0.807 and ROC-AUC of 0.857 at a privacy budget of ε = 8. Under a PGD attack of 0.5 °C per feature, its ROC-AUC was 0.613 against 0.446 for an undefended network, and membership-inference AUC fell from 0.546 to 0.503. SHAP placed the strongest evidence in the medial and lateral plantar angiosomes, which were the localized region for 70.5% of diabetic subjects. The results show that robustness and privacy can be added to thermographic screening at a measurable cost in clean accuracy, and that angiosome-level explanations give clinicians a location to examine.

Keywords: *Diabetic foot, infrared thermography, explainable AI, SHAP, adversarial robustness, differential privacy, angiosome*

1. Introduction

About one in three people with diabetes will develop a foot ulcer in their lifetime, and ulcers precede most diabetes-related lower-limb amputations [1], [2]. India has one of the world's largest diabetic populations [3], but routine foot examination outside specialist clinics is irregular. Early signs of ulceration — local inflammation, impaired blood flow and neuropathy — change plantar skin temperature before a wound is visible, and infrared thermography measures this temperature without contact, at low cost [4], [5].

Machine learning on plantar thermograms has shown promise for separating diabetic from healthy feet and for flagging regions at risk [6], [7]. Accuracy alone, however, is not enough for a screening tool. First, clinicians need to know why and where a model raises an alarm. Second, temperature readings carry sensor noise and calibration drift, and a model whose output flips under small changes cannot be trusted; deliberate adversarial perturbations are an established threat to medical AI [8]. Third, thermograms and their derived features are patient health data, and trained models can reveal whether a given patient was in the training set [9], [10].

This paper addresses all three together. Its contributions are:

- A numerical feature pipeline that summarizes each foot and each of its four plantar angiosomes, plus left–right asymmetry, from raw temperature maps.
- A detector trained with PGD adversarial training (budget in °C) and DP-SGD (formal (ε, δ) guarantee), compared with classical and undefended baselines.

- SHAP explanations aggregated per angiosome, giving a per-patient localization map, with a deletion test of explanation faithfulness.
- A systematic study of the trade-off between clean accuracy, robustness under FGSM/PGD attacks, privacy budget and membership-inference leakage.

2. Related Work

Thermography for diabetic foot: Contralateral temperature differences above about 2.2 °C are a recognized warning sign for ulceration [4], [5]. Hernández-Contreras et al. released a plantar thermogram database of diabetic and control subjects segmented into the four plantar angiosomes — medial calcaneal (MCA), lateral calcaneal (LCA), medial plantar (MPA) and lateral plantar (LPA) — and proposed a thermal change index [11]. Subsequent studies applied classical classifiers and deep networks to this database for diabetic foot classification [6], [7], [17]. These works report accuracy but do not test robustness or privacy, and explanations, where given, are not tied to anatomy.

Explainable AI: SHAP assigns each feature an additive contribution to a prediction based on Shapley values [12]. Faithfulness can be checked by deleting the highest-ranked features and measuring the drop in performance [13].

Adversarial robustness: FGSM and PGD craft worst-case perturbations within a norm ball; adversarial training with PGD is the standard defence [14], [15]. Finlayson et al. argued that medical AI is especially exposed to such attacks [8].

Privacy-preserving learning: DP-SGD clips per-sample gradients and adds Gaussian noise, bounding what a model reveals about any training record [9]. Membership-inference attacks test this leakage empirically [10].

Gap: To our knowledge, no prior work on diabetic foot thermography combines anatomically grounded explanations, adversarial robustness and formal privacy in one framework.

3. Dataset and Features

The Plantar Thermogram Database [11] contains thermograms of 167 subjects: 122 with diabetes mellitus (DM group) and 45 without (control group). For each subject it provides per-pixel temperature maps (CSV, °C) of the left and right plantar surface, and the same maps segmented into four angiosomes. The dataset is open access on IEEE DataPort.

Feature extraction. For each foot (left, right) and each region (whole foot, LCA, LPA, MCA, MPA), five statistics are computed over the non-zero pixels: mean, standard deviation, 10th and 90th percentile, and maximum. Contralateral asymmetry |mean_left
– mean_right| is added for the whole foot and each angiosome. This gives 55 features, all in °C. Demographic variables (age, sex, BMI) are deliberately excluded so that the model relies on thermal information only.

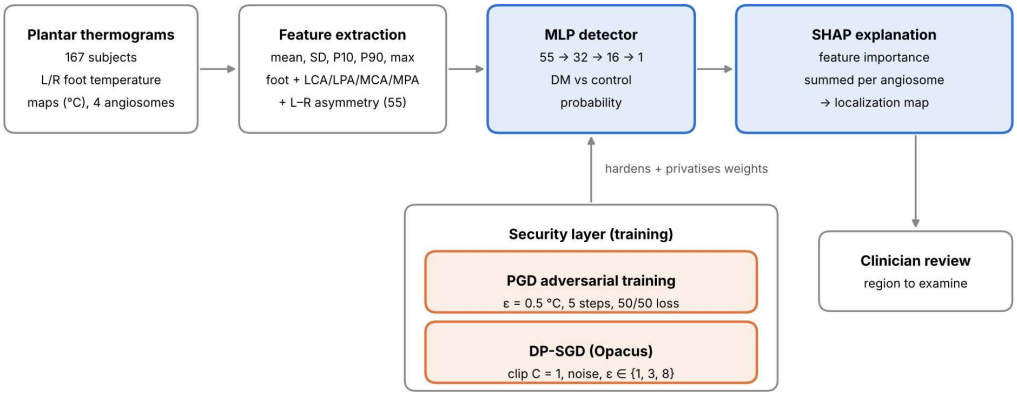
Table 1. Dataset summary

Group	Subjects	Age, mean ± SD (years)	Female / Male
Control	45	27.8 ± 8.2	16 / 29
Diabetes (DM)	122	56.0 ± 10.6	89 / 33
Total	167	—	105 / 62

4. Proposed Methodology

Fig. 1 shows the framework. Features are standardized using training-fold statistics only. The detector is a multilayer perceptron (MLP) with two hidden layers (32 and 16 units, ReLU, dropout 0.2) and a sigmoid output; class imbalance is handled with a positive-class weight in the binary cross-entropy loss.

Fig. 1. Proposed framework. Adversarial training and DP-SGD act during training; at inference SHAP explains each prediction and is aggregated per angiosome.



4.1. Adversarial robustness

Perturbations are bounded per feature in physical units: an attack of budget ε °C may shift each temperature feature by at most ε °C ($\epsilon/\sigma j$ in standardized units). FGSM takes one signed-gradient step; PGD takes T steps of size η with projection Π onto the budget:

$$x^{t+1} = \Pi_{\epsilon}\left(x^{t} + \eta\, \mathrm{sign}(\nabla_{x}\mathcal{L}(f(x^{t}), y))\right) \quad (1)$$

Adversarial training minimizes an equal mix of clean and PGD loss (5 steps, ε = 0.5 °C):

$$\mathcal{L}_{adv} = 0.5\,\mathcal{L}(f(x), y) + 0.5\,\mathcal{L}(f(x_{PGD}), y) \quad (2)$$

4.2. Privacy Preserving training

DP-SGD [9] clips each per-sample gradient to norm C = 1, adds Gaussian noise of scale σC to the batch sum, and tracks the cumulative privacy loss with a Rényi accountant:

$$\tilde{g} = \frac{1}{B}\left(\sum_{i=1}^{B} g_i \min\left(1, \frac{C}{\|g_i\|_2}\right) + \mathcal{N}(0, \sigma^2 C^2 I)\right) \quad (3)$$

The noise multiplier σ is set so that training ends at a target budget □ □ {1, 3, 8} with δ = 1/N. Adversarial examples for each batch are crafted on a copy of the current weights, so the attack does not consume privacy budget beyond the gradient step itself. Implementation uses PyTorch and Opacus [16].

4.3. Explanation and angiosome localization

SHAP values Ir are computed with KernelSHAP [12] against a 15-centroid k-means background of the training fold. The importance of a region r on a given foot is the sum of |Ir| over the five features of that region:

$$I_r = \sum_{j \in r} |\phi_j| \tag{4}$$

The angiosome with the largest Ir is reported as the localized region for that subject, and Ir is drawn on the angiosome masks to form a localization map. Faithfulness is tested by replacing the top-k features (global SHAP rank) with their training mean and measuring the ROC-AUC drop, compared with replacing k random features.

5. Experimental Setup

Protocol. Stratified 5-fold cross-validation repeated with 5 random seeds (25 runs per model). All scaling, background sets and thresholds use training folds only. Results are mean ± standard deviation.

Compared models. Logistic regression, SVM (RBF kernel) and random forest (500 trees) as classical baselines; MLP baseline; MLP + adversarial training; MLP + DP-SGD (ε = 8); and the proposed MLP + adversarial training + DP-SGD at ε = 8, 3 and 1. The MLP uses Adam (learning rate 3 × 10−3, weight decay 10−4), batch size 32 and 150 epochs.

Metrics. Accuracy, precision, recall, F1-score (decision threshold 0.5) and ROC-AUC. Robustness: the same metrics under FGSM and PGD-20 at ε = 0.25, 0.5 and 1.0 °C. Privacy: final ε and the ROC-AUC of a loss-threshold membership-inference attack (MIA) that separates training from test subjects; 0.5 means no measurable leakage. Explainability: deletion ROC-AUC and the distribution of localized angiosomes.

6. Results and Discussion

6.1 Detection performance

Table 2. Detection performance (mean ± SD, 25 runs)

Model	Accuracy	Precision	Recall	F1-score	ROC-AUC
Logistic Regression	0.850 ± 0.056	0.948 ± 0.047	0.842 ± 0.064	0.891 ± 0.043	0.933 ± 0.045
SVM (RBF)	0.867 ± 0.059	0.914 ± 0.050	0.905 ± 0.065	0.908 ± 0.042	0.933 ± 0.044
Random Forest	0.842 ± 0.066	0.920 ± 0.054	0.860 ± 0.072	0.887 ± 0.049	0.921 ± 0.050
MLP baseline	0.876 ± 0.066	0.940 ± 0.053	0.890 ± 0.067	0.912 ± 0.047	0.934 ± 0.047
MLP + Adv	0.799 ± 0.051	0.978 ± 0.032	0.743 ± 0.071	0.842 ± 0.046	0.911 ± 0.043
MLP + DP (e=8)	0.828 ± 0.074	0.933 ± 0.050	0.828 ± 0.099	0.873 ± 0.061	0.918 ± 0.052
Proposed (Adv + DP, e=8)	0.758 ± 0.057	0.954 ± 0.043	0.705 ± 0.081	0.807 ± 0.055	0.857 ± 0.063

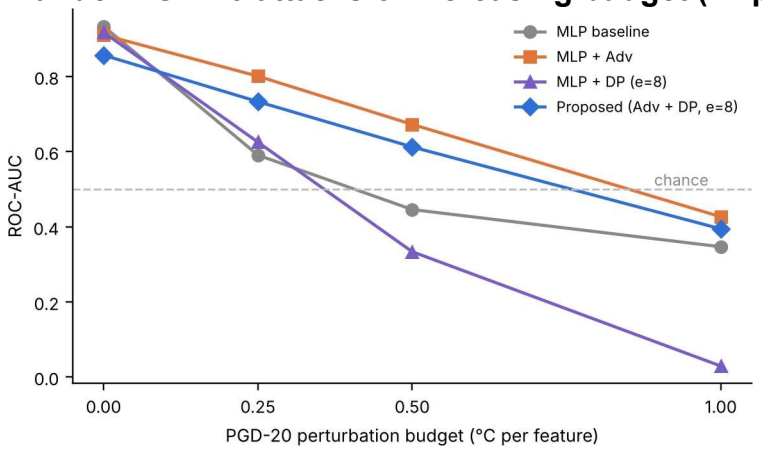
Thermal features alone separate diabetic from control subjects well: the undefended MLP reached an ROC-AUC of 0.934 and an F1-score of 0.912, on par with logistic regression (0.933) and SVM (0.933) and slightly above random forest (0.921). Each security layer costs some clean performance. Adversarial training lowered ROC-AUC to 0.911 and DP-SGD to 0.918; combined, the proposed model reached 0.857. The drop comes mainly from recall (0.890 → 0.705), while precision stayed high (0.954): the hardened model flags fewer subjects but is rarely wrong when it does.

6.2 Robustness under attack

Table 3. ROC-AUC under FGSM and PGD-20 attacks (budget in °C per feature; mean of 25 runs)

Model	clean	FGSM 0.25C	FGSM 0.5C	FGSM 1.0C	PGD 0.25C	PGD 0.5C	PGD 1.0C
MLP baseline	0.934	0.622	0.525	0.394	0.591	0.446	0.347
MLP + Adv	0.911	0.812	0.712	0.546	0.801	0.673	0.427
MLP + DP (e=8)	0.918	0.649	0.390	0.080	0.626	0.334	0.029
Proposed (Adv + DP, e=8)	0.857	0.750	0.672	0.546	0.733	0.613	0.394

Fig. 2. ROC-AUC under PGD-20 attacks of increasing budget (°C per feature).



The undefended MLP is fragile. A PGD perturbation of only 0.25 °C per feature — within the calibration error of many thermal cameras — cut its ROC-AUC from 0.934 to 0.591, and at 0.5 °C it fell below chance (0.446). DP-SGD alone did not help and was the most fragile model at larger budgets (0.029 at 1.0 °C). Adversarial training gave the strongest robustness (0.801 at 0.25°C, 0.673 at 0.5 °C). The proposed model kept most of this gain while also being private: 0.733 at 0.25 °C and 0.613 at 0.5°C, i.e. 0.142 and 0.167 above the undefended network. At 1.0 °C, beyond the training budget, every model was near or below chance, so robustness holds only up to about the budget used in training. F1-scores followed the same order: under PGD at 0.5°C the proposed model kept an F1-score of 0.683, against 0.645 for the undefended MLP and 0.509 for DP-SGD alone.

6.3 Privacy

Table 4. Privacy budget and membership-inference leakage (mean ± SD, 25 runs)

Model	DP budget e	Clean AUC	Clean F1	MIA AUC
MLP baseline	∞	0.934 ± 0.047	0.912 ± 0.047	0.546 ± 0.043
MLP + Adv	∞	0.911 ± 0.043	0.842 ± 0.046	0.526 ± 0.042
MLP + DP (e=8)	8.00	0.918 ± 0.052	0.873 ± 0.061	0.511 ± 0.048
Proposed (Adv + DP, e=8)	8.00	0.857 ± 0.063	0.807 ± 0.055	0.503 ± 0.046
Proposed (Adv + DP, e=3)	3.00	0.858 ± 0.064	0.813 ± 0.053	0.503 ± 0.046
Proposed (Adv + DP, e=1)	1.00	0.858 ± 0.071	0.796 ± 0.094	0.501 ± 0.043

The membership-inference attack separated training from test subjects with an AUC of 0.546 for the undefended MLP. Adver-sarial training reduced this to 0.526, and DP-SGD brought it to 0.501–0.511,

which is indistinguishable from chance given the run-to-run spread (±0.04–0.05). Tightening the budget from ε = 8 to ε = 1 did not lower the proposed model's ROC-AUC (0.857 vs 0.858), so the strongest guarantee (ε = 1) can be used at no further cost on this dataset. The leakage of the undefended model is modest, reflecting the small network and strong regularization; the value of DP-SGD here is the formal guarantee rather than a large empirical reduction.

6.4 Explanation and Localization

Fig. 3. Fifteen most important features of the proposed model by mean |SHAP value| (25 runs).

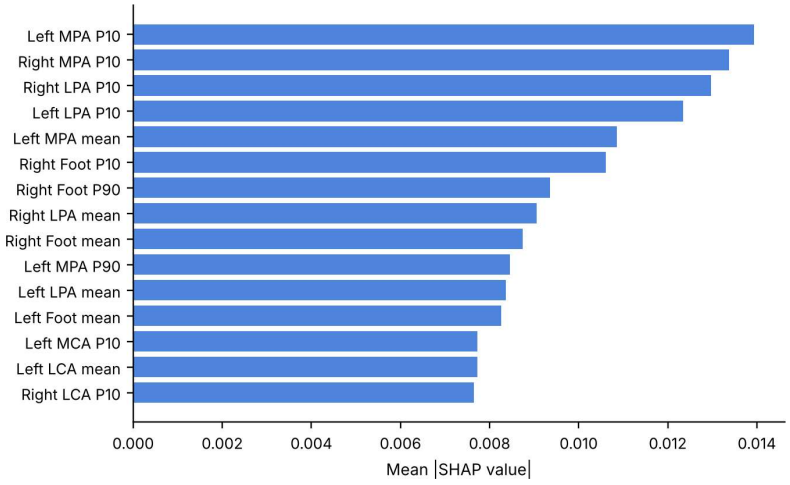
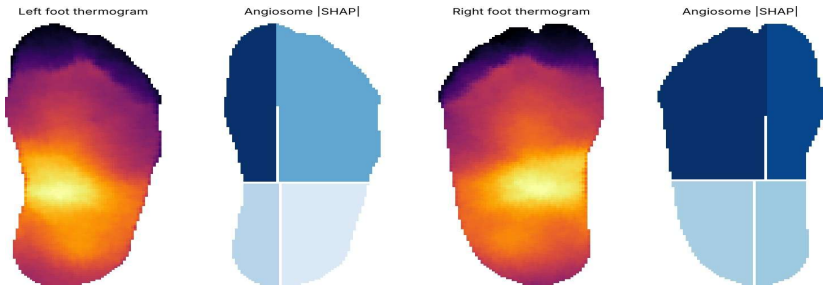


Table 5. Share of total |SHAP| importance per region (proposed model) and most important angiosome per DM subject

Region	Left foot (%)	Right foot (%)	DM subjects localized here
Medial plantar (MPA)	13.6	10.9	46 (37.7%)
Lateral plantar (LPA)	11.1	11.4	40 (32.8%)
Lateral calcaneal (LCA)	7.5	8.4	24 (19.7%)
Medial calcaneal (MCA)	6.9	6.7	12 (9.8%)
Whole foot	8.8	11.8	—
L–R asymmetry (all five)	3.0 (combined)		—

Fig. 4. Localization for one DM subject: plantar thermograms and per-angiosome SHAP importance (darker = more important).



The 10th-percentile and mean temperatures of the medial and lateral plantar angiosomes (MPA, LPA) of both feet were the most important features (Fig. 3). Grouped by region (Table 5), the plantar (forefoot–midfoot) angiosomes carried 47.0% of all impor-tance and the calcaneal (heel) angiosomes 29.5%; left–right asymmetry features contributed only 3.0%. Higher temperatures in these regions pushed predictions towards the diabetic class (correlation between feature value and SHAP value 0.57–0.67); the 10th-percentile plantar temperature was about 4.2 °C higher in the DM group than in controls (e.g. left MPA 28.95 vs 24.66 °C), consistent with the elevated plantar temperatures reported for this database [11]. For each diabetic subject, the angiosome with the largest summed |SHAP| was taken as its localized region: MPA for 37.7% and LPA for 32.8% of subjects (Fig. 4 shows one example).

The deletion test did not confirm faithfulness. Replacing the top 5, 10 and 20 SHAP-ranked features with their training mean did not lower ROC-AUC (0.867, 0.878 and 0.891 vs 0.857), and replacing 10 random features gave 0.856. The thermal features are highly correlated — every region's mean, percentiles and maximum move together — so the model can recover the same information from the features that remain. Explanations should therefore be read at the region level rather than for single features, and a correlation-aware attribution method is needed in future work.

6. Limitations

The control group is much younger than the DM group (27.8 vs 56.0 years), and age alone separates the two groups with a cross-validated ROC-AUC of 0.97. Although age was excluded from the features, plantar temperature also changes with age, so part of the measured separation may reflect age rather than diabetes. The dataset has 167 subjects from a single site, the label is diabetes status rather than ulcer outcome, and the angiosome localization has not yet been compared with clinical findings. DP guarantees are per subject here, since each subject contributes one record.

7.Conclusion and Future Work

This paper presented a secure and explainable framework for diabetic foot thermography that combines an MLP detector, PGD adversarial training, DP-SGD and angiosome-level SHAP localization. On the Plantar Thermogram Database, thermal features alone detected diabetic feet with an ROC-AUC of

0.934. Adding adversarial training and DP-SGD lowered clean ROC-AUC to 0.857 but raised ROC-AUC under a 0.5 °C PGD attack from 0.446 to 0.613 and reduced membership-inference leakage to chance level, with $\varepsilon = 1$ costing no more than $\varepsilon = 8$. SHAP localized most diabetic subjects to the medial and lateral plantar angiosomes, although a deletion test showed that single-feature attributions are not reliable because of feature correlation. Future work will collect age-matched thermograms with ulcer follow-up from additional clinical sites, validate localization against clinician examination, and extend the framework to thermal images with convolutional detectors.

References

1. *D. G. Armstrong, A. J. M. Boulton, and S. A. Bus, "Diabetic foot ulcers and their recurrence," N. Engl. J. Med., vol. 376, no. 24, pp. 2367–2375, 2017.*
2. *M. H. Yap et al., "Deep learning in diabetic foot ulcers detection: A comprehensive evaluation," Comput. Biol. Med., vol. 135, p. 104596, 2021.*
3. *International Diabetes Federation, IDF Diabetes Atlas, 10th ed. Brussels, Belgium: IDF, 2021.*
4. *D. G. Armstrong et al., "Skin temperature monitoring reduces the risk for diabetic foot ulceration in high-risk patients," Am. J. Med., vol. 120, no. 12, pp. 1042–1046, 2007.*
5. *L. A. Lavery et al., "Home monitoring of foot skin temperatures to prevent ulceration," Diabetes Care, vol. 27, no. 11, pp. 2642–2647, 2004.*
6. *A. Khandakar et al., "Thermal change index-based diabetic foot thermogram image classification using machine learning techniques," Sensors, vol. 22, no. 5, p. 1793, 2022.*
7. *I. Cruz-Vega et al., "Deep learning classification for diabetic foot thermograms," Sensors, vol. 20, no. 6, p. 1762, 2020.*
8. *S. G. Finlayson et al., "Adversarial attacks on medical machine learning," Science, vol. 363, no. 6433, pp. 1287–1289, 2019.*
9. *M. Abadi et al., "Deep learning with differential privacy," in Proc. ACM CCS, 2016, pp. 308–318.*
10. *R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in Proc. IEEE S&P, 2017, pp. 3–18.*
11. *D. Hernández-Contreras, H. Peregrina-Barreto, J. Rangel-Magdaleno, and F. Renero-Carrillo, "Plantar thermogram database for the study of diabetic foot complications," IEEE Access, vol. 7, pp. 161296–161307, 2019; dataset: IEEE DataPort, doi: 10.21227/tm4t-9n15.*
12. *S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Proc. NeurIPS, 2017, pp. 4765–4774.*
13. *V. Petsiuk, A. Das, and K. Saenko, "RISE: Randomized input sampling for explanation of black-box models," in Proc. BMVC, 2018.*
14. *I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in Proc. ICLR, 2015.*
15. *A. Madry et al., "Towards deep learning models resistant to adversarial attacks," in Proc. ICLR, 2018.*
16. *A. Yousefpour et al., "Opacus: User-friendly differential privacy library in*

PyTorch," arXiv:2109.12298, 2021.

17. *A. Khandakar et al., "A machine learning model for early detection of diabetic foot using thermogram images," Computer in Biol. Med.*, vol. 137, p. 104838, 2021.