IEEE
INSTRUMENTATION
& MEASUREMENT
SOCIETY®

Received 28 June 2022; revised 14 August 2022; accepted 18 August 2022. Date of publication 1 September 2022; date of current version 6 October 2022.
The review of this article was arranged by Associate Editor Jacob Scharcanski.

Digital Object Identifier 10.1109/OJIM.2022.3203452

Convex Factorization Embedding Thermography for Breast Cancer Diagnostic

NICOLLE VIGIL [1], BEHROUZ MOVAHHED NOURI [2,3], HENRIQUE C. FERNANDES [4], CLEMENTE IBARRA-CASTANEDO [5], XAVIER P. V. MALDAGUE [5], AND BARDIA YOUSEFI [1]

[1] Fischell Department of Bioengineering, University of Maryland, College Park, MD 20742, USA
[2] Department of Electrical and Computer Engineering, George Washington University, Washington, DC 20052, USA
[3] Optelligence Company, Austin, TX 78701, USA
[4] Faculty of Computing, Federal University of Uberlandia, Uberlandia 38408, Brazil
[5] Department of Electrical and Computer Engineering, Laval University, Québec, QC G1V 0A6, Canada
CORRESPONDING AUTHOR: B. YOUSEFI (e-mail: byousefi@umd.edu)

ABSTRACT Thermographic imaging has proven to be effective for the early detection of breast cancer and with clinical breast examination (CBE). There are many matrix factorization methods developed for computational thermography that can be used to extract thermal variations across the acquisition time. These methods are often used to summarize thermographic sequences and simultaneously highlight predominant thermal patterns. Finding a single predominant infrared image capturing the prevalent patterns of changes remains a challenging task in the field. This study presents the applications of convex factor analysis combined with the bell-curve membership function embedding approach to tackle this task and generate one image to represent the entire sequence. This low-dimensional (LD) representation of a thermal sequence was then used to extract thermomics and train tuned hyperparameters random forest model for early breast cancer diagnosis. A comparative analysis of different embedding methods and factorization approaches is also provided. The results of the proposed method combining clinical information, and demographics yield 78.9% (75.7% and 85.9%), while the convex-nonnegative matrix factorization (NMF) alone gave 76.9% (73.7% and 86.1%). The result of the proposed method suggests that the embedding can help preserve important thermal patterns, which significantly aid CBE and early detection of breast cancer.

INDEX TERMS Breast cancer diagnosis, data dimensionality reduction, embedding, factor analysis, thermomics.

I. INTRODUCTION

THERMOMICS, imaging thermal features, have been proven to be effective in warning physicians about early breast cancer as the second cause of death in women [1]. Thermography is proposed to be used for a clinical breast examination (CBE) and before performing mammography acquisition, which can provide information about any potential abnormality in the patients [2], [3]. Thermographic imaging works due to an increase in the vasodilation and angiogenesis blood vessel formation in the breast area because of irregular lesions. Such endocrine alterations because of the breast lesions change the thermal profile representing vascularization for supplying oxygen and nutrients to lesions [3], [4]. An infrared camera can capture such changes, which ultimately leads to finding abnormalities (see Fig. 1). Several studies substantiated the importance of thermography in sensing hypervascularity in nonpalpable breast cancer [4]. This can be used as a potential biomarker for an early finding of breast cancer with the CBE and before mammography.

But the biggest challenge here is to capture such irregularities using infrared technology and the transition between raw infrared sequences to thermal patterns. Low-rank matrix approximation methods by selecting the predominant basis of the decomposed eigenvector matrix were being wildly applied to extract the predominant images representing the entire thermal stream. Some well-known techniques are

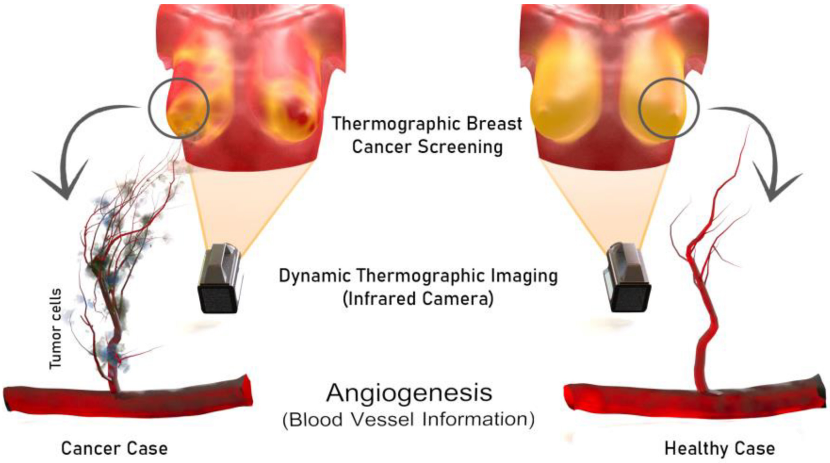

FIGURE 1. Schematic of the application of infrared thermography to detect heterogeneous thermal patterns.

TABLE 1. Table of notations.

NOTATION	DESCRIPTION	NOTATION	DESCRIPTION	NOTATION	DESCRIPTION	NOTATION	DESCRIPTION
$\boldsymbol{s}$	Vector size of the thermal image	β	Basis tensor	$\mathbf{w}_{\boldsymbol{\ell}}$	Weight of data points	$\boldsymbol{\zeta}_{\mathbf{i}}$	samples stay in every block
$\mathbf{m,n}$	Height and width of image	$\mathbf{H}$	Coefficient matrix	$\boldsymbol{\Phi}$	Gaussian Embedded bases	$\dot{\boldsymbol{\Phi}}$	Bell embedding bases
$\boldsymbol{\tau}$	temporal size	$\|\|B\|\|^2$	ℓ_2-norm of B	$\boldsymbol{\eta}_{\mathbf{i}}$	Gaussian embedding	b	Obituary attention coefficient
$\mathbf{X} \in \mathbb{R}^{\pm}_{\mathbf{MN}\times\boldsymbol{\tau}}$	vectorized thermal images, heat matrix	$\mathbf{X}^{\pm}$	normalized X	μ	means of thermal reference		
$\mathbf{B} \in \mathbb{R}^{\pm}_{\boldsymbol{MN}\times\boldsymbol{\tau}}$	Bases matrix	$\mathbf{B}^{\pm}$	Bases (- or +)	σ	standard deviation		

principal component analysis (PCT) [5], [6], nonnegative matrix factorization (NMF) [7], Fixed Eignvector analysis [8], incremental PCT [5], Sparse factorization [10], *t*-distributed stochastic neighborhood embedding (tSNE) [11], candid covariance-free incremental principal component thermography (CCIPCT) [12], sparse PCT [13], [14], semi NMF [15], [16], [17], sparse NMF [18], convex NMF [19], deep NMF [20], and deep learning convex [21] in thermography.

An Additional difficulty in extracting low-rank matrix representation is selecting the appropriate basis corresponding to the maximum variance of the thermal pattern. The aforementioned problem is tackled by embedding [20] and here this study shows the application of embedding in convex factorization analysis in thermography and to extract the low-dimensional (LD) representation of the thermal sequence in the form of the most predominant basis combined with embedding [20] and use it to extract thermomics and training a classifier for early diagnosis of breast abnormality. The proposed Bell-curve membership function for embedding better represents the normal distribution than the previously proposed Gaussian embedding and that may increase the better representation of the thermal patterns and ultimately more appropriate biomarker.

Our results were compared to the gold standard breast cancer screening modality, mammography images, and breast biopsy to ensure the accuracy of the model. This study shows a modification of basis embedding in thermography [18], [20] and another example of the reliability of thermomics for the early detection of breast cancer. Table 1 shows the mathematical notations used in this article.

II. METHOD

Tracking the heterogeneous thermal patterns in infrared imaging has been demonstrated to be helpful in the early diagnosis of breast cancer. Here, we present a methodology for the detection of thermal patterns through factorization analysis and a novel embedding (Fig. 2).

A. RELATED WORKS

The heterogeneous thermal patterns are an indication of an abnormal thermal property of the parenchymal tissues [3], [4], [11], [18], [19], [20], [21]. This can be computed using approaches developed for computational thermography to extract the predominant basis across thermal dimensions and also represent temporal pattern variations in the infrared images, i.e., different eigendecomposition and matrix factorization techniques [5], [6], [7], [8], [9], [10], [11], [12],

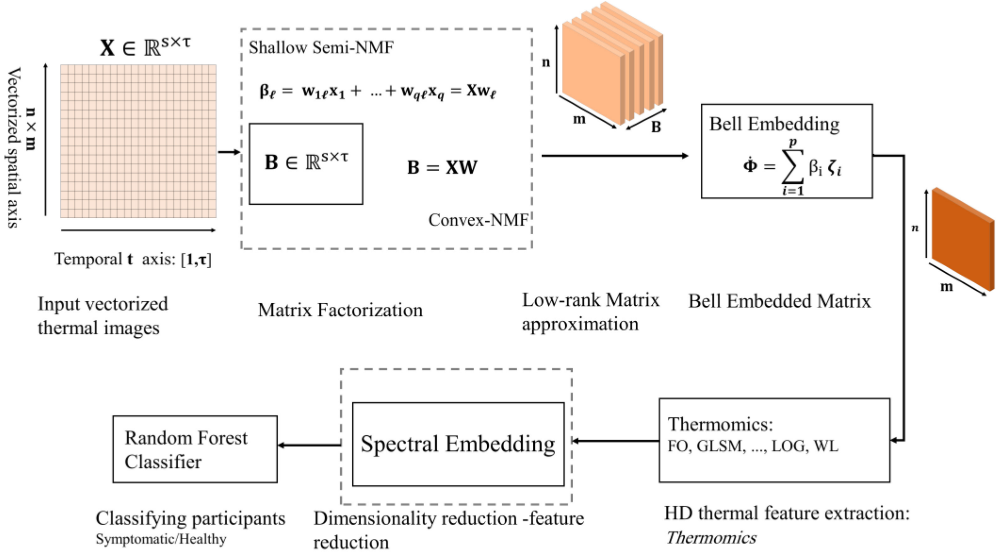
$\mathbf{X} \in \mathbb{R}^{s\times\tau}$
Vectorized spatial axis
n × m
Temporal t axis: [1,τ]
Input vectorized thermal images
Shallow Semi-NMF
$\beta_\ell = w_{1\ell}x_1 + \ldots + w_{q\ell}x_q = Xw_\ell$
$\mathbf{B} \in \mathbb{R}^{s\times\tau}$
B = XW
Convex-NMF
Matrix Factorization
n
m
B
Low-rank Matrix approximation
Bell Embedding
$\dot{\Phi} = \sum_{i=1}^{p} \beta_i \zeta_i$
Bell Embedded Matrix
n
m
Thermomics:
FO, GLSM, ..., LOG, WL
HD thermal feature extraction: Thermomics
Spectral Embedding
Dimensionality reduction -feature reduction
Random Forest Classifier
Classifying participants
Symptomatic/Healthy

FIGURE 2. Workflow of the proposed approach using Convex-NMF with low rank Bell embedding method is presented.

[13], [14], [15], [16], [17], [18], [19], [20]. Principal component analysis (PCA) in thermography or PCT [5], [6], provides an LD representation of thermal sequence using covariance calculation of heat matrix. This can be executed by singular value decomposition (SVD), which shared many properties with CCIPCT [9], [12] and sparse PCT [13], [14]. CCIPCT and Sparse PCT are the modifications of PCT while converting that algorithm into incremental without covariance calculation approach and with additional regularization parameters to increase the speed and robustness of the model, respectively. If we restrict the bases and coefficients of PCA with nonnegative constraints, we can have NMF [7], [16], [18] and if we lose these constraints, we can modify the NMF to semiNMF [17], [19], and Convex-NMF [18]. Sparse NMF [17], [18], [19] is comparable in terms of methodology to Sparse PCT [13], [14] to increase the robustness of the decomposition through regularization terms. The deep semi NMF decomposes the basis matrix into many hidden bases, which delivers sparse representation. Deep basis layers also can be trained while preserving different thermal patterns [20].

These methods are using the pairwise distance between temporal points in thermal sequences and try to preserve them while they are transferred onto lower dimensional space. T-distributed stochastic neighbor embedding is also used in thermography, but it replaces Euclidean distance between pairwise points in thermal sequence with a stochastic measure of similarity and tries to minimize the Kullback–Leibler divergence between the probability of two-point centered Gaussian distributions [11]. The challenge here is to select predominant bases with any of the aforementioned algorithms, which we address this problem with embedding and a comparison analysis on previously developed models to ensure the reliability of the diagnostic system.

B. CONVEX FACTORIZATION IN THERMOGRAPHY

The NMF represents linearly with the nonnegative approximation of data. Heat matrix, $\mathbf{X} \in \mathbb{R}^{s\times\tau}$, $s = nm$, where N and M are the spatial resolutions of thermal images. It gives a linear representation of bases to construct a heat matrix, which is very similar to PCA with nonnegative constraints. It decomposes the stacked vectorized infrared images to a set of vectors $\boldsymbol{B} = [\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_\tau]$ and $\boldsymbol{B} \in \mathbb{R}^{s\times\tau}$, which is a linear data approximation

$$\begin{aligned} &\boldsymbol{x_i} \approx \mathbf{B}\boldsymbol{\alpha_i} \quad \boldsymbol{i} = \mathbf{1}, \ldots, \boldsymbol{\tau} \\ &\text{s.t.} \quad \boldsymbol{\alpha_i} \geq \mathbf{0} \end{aligned} \tag{1}$$

where $\boldsymbol{\alpha} \subseteq \boldsymbol{H}$ and $\boldsymbol{\alpha_i} \in \mathbb{R}^{+}_{s\times 1}$ called the linear combination coefficient. The equation for the data's full element can be given in the matrix approximation format

$$\begin{aligned} &\boldsymbol{x_i} \approx \mathbf{B}\boldsymbol{\alpha_i} \quad \boldsymbol{i} = \mathbf{1}, \ldots, \boldsymbol{\tau} \\ &\text{s.t.} \quad \mathbf{B} \geq \mathbf{0} \quad \mathbf{H} \geq \mathbf{0}. \end{aligned}$$

Solving the aforementioned problem includes computation of the squared error or $\boldsymbol{\ell_2}$ equation that delivers the Euclidean distance [22] and loss to be followed for the maximum similarity of the bases to original images

$$\begin{aligned} &\min_{\mathbf{B},\mathbf{H}} \; f(\mathbf{B}, \mathbf{H}) = \|\mathbf{X} - \mathbf{BH}\|^2 \\ &\text{s.t.} \quad \boldsymbol{B} \geq \mathbf{0}, \quad \mathbf{H} \geq \mathbf{0}. \end{aligned} \tag{2}$$

NMF limits matrices **X**, **B**, and **H** to be nonnegative. **B** represents predominant thermal patterns obtained from **X**. But when the data matrix is unconstrained, it converts to Semi-NMF, in which **H** is constrained to be nonnegative but there is no limit for matrix **B**. This has duality with the k-means clustering.

In NMF and Semi-NMF, there are no constraints for basis vectors **B**, so it can be pertinent to enforce the constraint of basis vectors within the column space of **X**

$$\begin{aligned} &\boldsymbol{\beta_\ell} = \mathbf{w}_{1\ell}\mathbf{x}_1 + \mathbf{w}_{2\ell}\mathbf{x}_2 + \cdots + \mathbf{w}_{q\ell}\mathbf{x}_q = \mathbf{X}\mathbf{w}_\ell \\ &\mathbf{B} = \mathbf{XW}. \end{aligned} \tag{3}$$

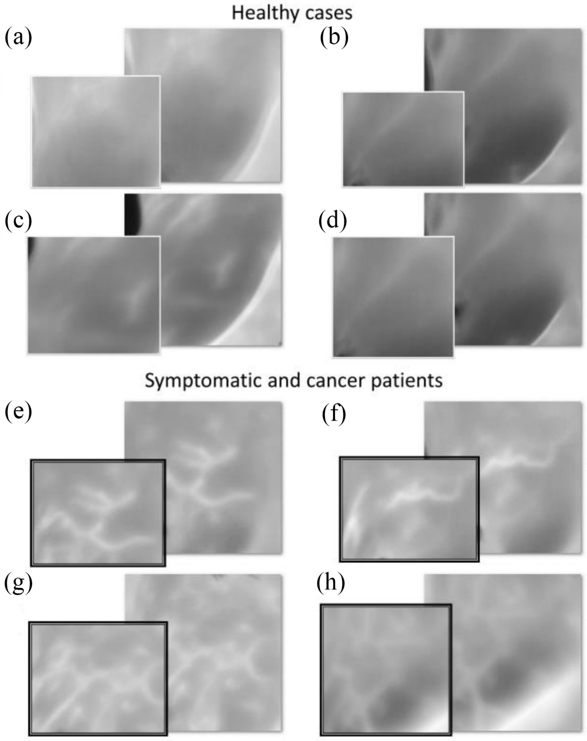

FIGURE 3. Eight examples of convex factor analysis. (a)–(d) Four healthy cases. (e)–(h) Four abnormal cases.

This focuses on the combination of columns of $\mathbf{X}$ which are needed to be convex. With that, we could get the columns β_ℓ as weighted summaries of data points. These columns induce the concept of the centroid, and this restricted form of $\mathbf{B}$ factor refers to as Convex-NMF [15]. Fig. 3 presents eight examples of convex factor analysis, four healthy cases [Fig. 3(a)–(d)], and four abnormal [Fig. 3(e)–(h)] cases of convex matrix factorization with the bell-shaped embedding function.

C. BELL-SHAPE EMBEDDING

Bases embedding is previously proposed for Deep-SemiNMF by combining multiple decomposed bases [20]. Using matrix factorization methods bases are generated and embedding helps combine k predominant tensors to reduce the dimensionality of presenting thermal images to one thermal image. Motivated from the same assumptions, we argue that the Convex-NMF converts higher temporal dimensionality to lower temporal representation and the obtained lower dimensional tensors can be considered as bases computed using matrix factorization approaches. A set of LD represented bases using Convex-NMF is shown as $\mathbf{B} = \{\beta_1, \beta_2, \ldots, \beta_p\}$, where $\mathbf{B} \in \mathbb{R}^{s\times\tau}$, $s =$ nm. We follow the definition of the membership function, Definition 2 (taken from [20]), to highlight and integrate the overall representation of these tensors.

Definition 2: The embedded LD representation, $\mathbf{\Phi}$, defines by aggregating membership calculated for $\boldsymbol{p}$ bases of $\boldsymbol{X}$, μ_p, multiply by the basis itself, β_i, and defined as

$$\mathbf{\Phi} = \sum_{i=1}^{p} \beta_i \eta_i \tag{4}$$

where $\boldsymbol{\eta}_{\mathbf{i}}$ is a membership of basis β_i and is defined by

$$\boldsymbol{\eta}_{\mathbf{i}} = \mathbf{e}^{\frac{\beta_i - \mu}{\sigma}}.$$

Let μ, σ mean (average) of thermal basis, and standard deviation of ith basis in the calculation. In this definition, $\mathbf{\Phi} \in \mathbb{R}^{s\times 1}$, $\boldsymbol{X} \in \mathbb{R}^{s\times\tau}$, and $p \ll \tau$. Here, we provide a derivation of this definition for bell curve Cauchy distribution as follows.

Definition 3: The embedded low-rank matrix approximation, $\dot{\mathbf{\Phi}}$, defined by aggregating membership calculated for $\boldsymbol{p}$ bases of $\boldsymbol{X}$, μ_p, multiply by the basis itself, β_i, and defined as

$$\dot{\mathbf{\Phi}} = \sum_{\boldsymbol{i}=\mathbf{1}}^{\boldsymbol{p}} \beta_i \; \boldsymbol{\zeta}_{\boldsymbol{i}} \tag{5}$$

where $\boldsymbol{\zeta}_{\boldsymbol{i}}$ is a membership of basis β_i and is defined by

$$\zeta_{\mathbf{i}} = \frac{1}{1 + \left|\frac{\beta_i - \mu}{\sigma}\right|^{2b}}$$

where b is an obituary coefficient and can increase the attention of the membership function ($b = 1$ in this article). $\boldsymbol{\zeta}_{\boldsymbol{i}}$ is a generalized bell curve (or Bell-shaped Function) and a direct generalization of the Cauchy distribution.

The insight behind applying to embed involves underlining the thermal variation in exponential order, which improves the thermal heterogeneity in the accumulated resulting image of the thermal sequence, and we previously named it the *avatar*.

D. THERMOMICS DIMENSIONALITY REDUCTION

The extraction of quantitative features from different medical imaging modalities is known as Radiomics, which undeniably enhances computer-aided decision, which is the diagnosis, in various cancer imaging analyses. With the help of various filters to translate this information and deliver analytical responses determined from such data. In infrared thermography, radiomic features are recognized as *thermomics* [17], [18], [19], [20] and are widely employed to diagnose breast cancer in an early stage before mammography. Heterogeneous thermomics are an indicator, or biomarker, for thermal patterns projecting vasodilation of abnormal breast tissues [15], [17], [18], [19], [20], [21], [22]. After generating avatars, we extract high-dimensional (HD) thermomics using the Pyradiomics library [23].

Here, we used the embedding thermal image and region of interest (ROI) to extract HD thermomics to qualify the thermal variations. We employed many features, and we reduce the dimensionality of thermomics using spectral embedding, to avoid overfitting the decision-making model. Then, we applied statistical analysis to show the connection between these features. In this study, we propose convex factorization embedding thermal for an early diagnostic system with thermomics produced by the Convex- NMF method for a sequence of dynamic thermographic images. The proposed Bell-shape membership function embedding helps emphasize bases extracted from the images as a competitor of Gaussian embedding.

TABLE 2. Clinical information and demographics of the breast cancer screening database using thermal imaging.

DMR - Database for Mastology Research		
Age	Median (±IQR)	60 (25,120)
Race	Caucasian	77 (37%)
	African	57 (27.4%)
	Pardo	72 (34.6%)
	Mulatto	1 (0.5%)
	Indigenous	1 (0.5%)
Diagnosis[1]	**Healthy**[2]	128 (61.5%)
	Symptomatic	80 (38.5%)
	Sick[3]	36 (17.3%)
Family history	Diabetes	52 (25%)
	Hypertensive	5 (2.4%)
	Leukemia	1 (0.5%)
	None	150 (72.1%)
Hormone therapy (HT)	Hormone replacement	38 (18.3%)
	None	170 (81.7%)

[1] This diagnosis performed with mammography as ground truth in this Dataset. [2] Healthy term is used as non-cancerous and asymptomatic patients. [3] We use the term "sick", which includes different types of breast cancer patients diagnosed by mammographic imaging.

III. RESULTS

Vasodilatation and blood formation were tracked using heterogeneous thermal patterns and has been tested on breast cancer screening datasets [24]. We generate the results of convex factorization embedding and compared them with the results from other low-rank matrix approximation algorithms to provide a comparative analysis.

A. BREAST CANCER STUDY DATA

Two hundred and eight participants, including healthy (without symptoms) or sick (cancer patients or symptomatic) cases were used to benchmark the proposed model. Cancer patients and symptomatic cases were used by CBE, mammography, and tissue biopsy but breast cancer cases and noncancerous but with symptoms were categorized as abnormal cases. The study group has a median age of 60 years, with Pardo, 77 (37%), 57 (27.4%) African, Caucasian 72 (34.6%), 1 (0.5%) indigenous, and 1 (0.5%) Mulatto women. 38 (18.3%) participants were undergoing hormone replacement and 52 (25%) cases had a history of diabetes in their families. All patients went through infrared acquisition with the following protocol: spatial resolution of images was 640×480 pixels. A FLIR thermal camera (model SC620) with a sensitivity of less than 0.04 °C range and captures a thermal range of −40 °C to 500 °C was utilized [24]. Table 2 shows the clinical information and demographics of the study cohort.

B. RESULTS OF CONVEX-NMF IN THERMOGRAPHY

Convex-NMF spanned thermal sequences with 23 dimensions to five LD thermal bases. Some examples of LD representation using Convex-NMF embedding are shown in Figs. 4 and 5. LD images significantly highlight heterogeneous thermal patterns in the breast area for more than 80 participants for breast cancer screening [sick and healthy with symptoms, Fig. 5(e)–(h)]. However, thermal patterns demonstrated by LD for the healthy participants represent more homogeneity [Fig. 5(a)–(d)].

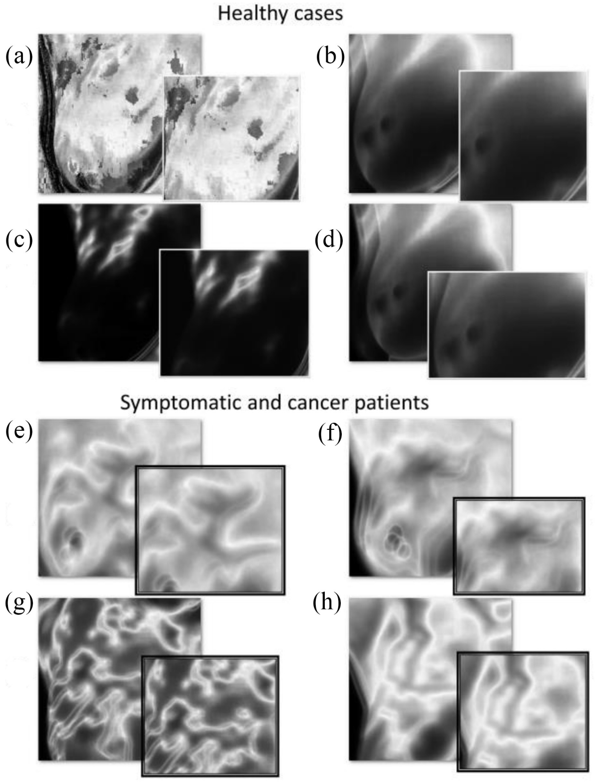

FIGURE 4. Eight examples of convex factorization Gaussian embedding (a)–(d) four healthy cases and (e)–(h) four abnormal cases.

C. RESULTS OF EMBEDDING

Five LD-represented images obtained by convex factorization methods were embedded using the proposed bell embedding approach. To establish the level of thermal heterogeneity in the breast area, we utilized the reference label, attached between the breasts of the participants as a reference point and to normalize the representation of images. Thermal heterogeneity is drastically heightened when applied to embed. This can considerably discriminate symptomatic and cancerous patients from healthy participants. Figs. 4 and 5 show a visual comparison between the heterogeneity of cases after two types of embedding, while Fig. 3 presents convex factorization without embedding.

D. THERMOMIC FEATURES AND CLASSIFICATION RESULTS

Three hundred and fifty four thermomics have been extracted from the breast areas, regions of interest-ROI, from embedded Convex-NMF generated avatar in four different feature categories: 1) first-order statistics; 2) texture; 3) intensity; and 4) spatiotemporal filtering features. To extract thermomic features, we used the Pyradiomics python library [23]. Then,

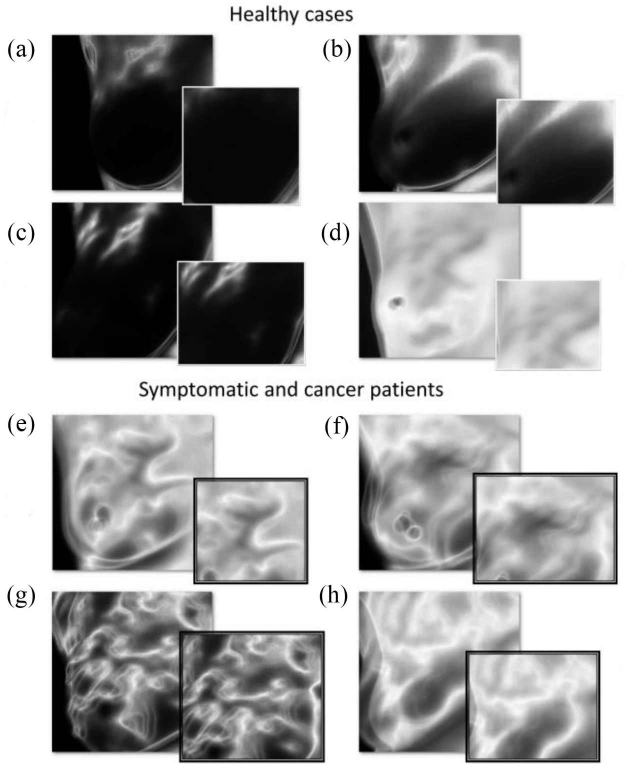

FIGURE 5. Eight examples, (a)–(d) four healthy cases and (e)–(h) four abnormal cases of convex matrix factorization with bell-shaped embedding function.

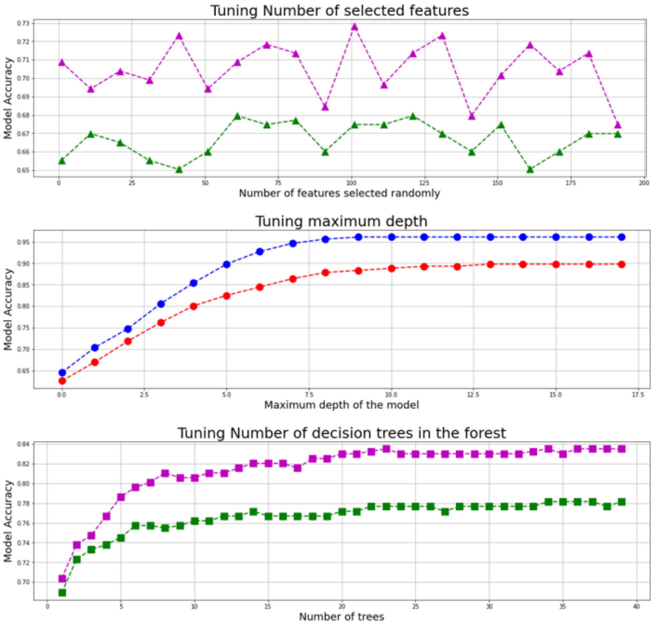

FIGURE 6. Hyperparameter tuning for the random forest, blue curves are representing deep radiomics and red curves show conventional radiomics fed to the model for the tuning using leave-one-out cross-validation.

features were concatenated in a matrix with 354 thermomics links to each vectorized thermal image. To lessen the collinearity of these HD attributes, spectral embedding reduces the dimensionality to seven features as an optimum number of thermomics.

To analyze the hypothesis that the LD embedded thermal heterogeneity can be utilized as a biomarker, we trained a random forest model with the generated LD thermomics for the study cohort. To extract the model's hyperparameters, we tuned the model for different scenarios, using a grid-search-motivated approach. The maximum depth, the random state in the tree, and the number of trees in the forest were optimized by changing the hyperparameters for access to our data using the leave-one-out cross-validation technique (see Fig. 6).

For the selected thermomics using different techniques, the classifier is fitted to the thermomic data and tested with multivariate analysis to classify patients using leave-one-out cross-validation. It resulted in 71.7% (69.1% and 75.1%) accuracy for clinical information and demographics and 77.6% (43.2% and 87.4%) for Convex NMF with Gaussian embedding, which were challenged by the Bell-shape embedding technique yielded to 76.9% (73.7% and 86.1%) accuracy and other approaches such as NMF, PCT, Deep SemiNMF, and CCIPCT showed a close range of accuracies with a median of 76.9% (Table 3). The highest accuracy belonged to Convex NMD Bell-embedding with Clinical information yielded to 78.9% (75.7% and 85.9%). Similarly, we established the embedding technique's strength by getting the accuracy of methods before and after embedding (for the same subset of thermomics) as a comparison. A *t*-test to Convex-NMF showed a significant statistical difference among the accuracies.

The computational process was performed using Python programming language. Convex-NMF and embedding methods showed considerably lower computational time than other approaches, in contrast to Sparse PCT and Spare NMF.

IV. DISCUSSION

In this study, we introduced another form of LD bases embedding in thermography to provide LD representation of thermal patterns for diagnostic purposes. The design of this study was for modifying the LD representation of the HD heat matrix and replacing matrix factorization techniques, i.e., [1], [2], [3], [4], [5], [6], [7], [8], [9], [14], and [15]. This study aims to improve the performance of HD infrared data LD representation. Previously, we proposed embedding [20] to generate LD thermal bases, whereas here Bell curve embedding enhances the process. Moreover, we focused on defining a new embedding procedure to unravel multiple LD representations and associated it with the earlier presented technique to see their outcomes for nonlinear data projection with convex matrix factorization models. Through this process, we demonstrated a likelihood to detect probable breast cancer patients by applying the proposed methods as a noninvasive, and cost-effective imaging procedure to aid CBE and physicians as an initial diagnostic tool prior to performing mammography or MRI.

Embedding showed substantial enhancements in classifying abnormal patients from healthy participants [20], changing the embedding function motivated by the Cauchy distribution provided a wider bell shape curve and showed

TABLE 3. Results of random forest classification in the leave-one-out cross-validation model.

Model	Factorization method	Embedding method	Classification Accuracy [2] (%)	Kappa coefficient (κ)
Random Forest	**Clinical**	-	71.7 (69.1, 75.1)	71.8 (69.3, 74.7)
	Convex-NMF	Gaussian	77.6 (43.2, 87.4)	78.1 (43.7, 87.9)
	CCIPCT	Bell	76.9 (45.7, 85.8)	77.6 (55.6, 87.3)
	PCT	Bell	76.9 (45.7, 85.6)	76.2 (43.7, 87.4)
	NMF	Bell	76.9 (45.8, 85.7)	76.5 (44.7, 87.5)
	Sparse NMF	Bell	77.9 (45.6, 87.4)	76.8 (43.6, 87.6)
	Sparse PCT	Bell	77.6 (45.7, 87.4)	77.2 (43.7, 87.6)
	Deep SemiNMF	Bell	76.4 (45.7, 84.8)	79.4 (55.6, 87.3)
	Convex-NMF	Bell	76.9 (73.7, 86.1)	78.7 (75.5, 86.8)
	Convex-NMF+Clinical	Bell	78.9 (75.7, 85.9)	78.3 (75.1, 85.7)

[1] The covariates used for the clinical and demographics were family history, age, and marital status. [2] Classification accuracy reported by median (±IQR) (Interquartile range-IQR). *t*-test calculated for each method versus maximal accuracy showed significant difference between convex-NMF and other accuracies.

considerable performance in highlighting thermal patterns (Figs. 4 and 5). Moreover, bell-shape embedding showed a slightly higher intensity and contrast profile of the integrated bases than the Gaussian approach (Fig. 4), which leads to the detection of more heterogeneous thermal patterns (Fig. 5). This might be due to the wider plateau on the top of the Gaussian curve in the bell-shaped function.

However, Sparse factorization approaches showed slightly higher accuracies, Table 3, due to additive sparsity constraints inducing more robustness in capturing thermal patterns. Similarly, multilayer bases and aggravating constraints in Deep-SemiNMF caused relatively lower accuracy for this approach. Convex-NMF carries some properties of NMF and with that bases behave like clustering [19], [20]. Convex-NMF performs clustering with sharper indicators. $\boldsymbol{\beta}_\ell$ is close to centroids in Kmeans. Bases are much more restricted than the original NMF, which has large effects on the subspace factorization. With respect to $\mathbf{min}_{\mathbf{B},\mathbf{H}} \boldsymbol{f}(\mathbf{B},\mathbf{H})$ in (2) implies getting larger residual values delivering more constraints leading more degradation of grouping thermal patterns. Embedding not only integrates the thermal bases but also enhances thermal properties (yet repetitious) in factorization methods compensating for larger constraints.

The proposed approach generates HD thermomics, which intensified the possibility of overfitting the random forest model, the *curse of dimensionality* [25], [26]. We reduced the dimensionality of thermomics by capturing the predominant features representing thermal patterns using spectral embedding, which increases the robustness of the method. Similar to other proposed approaches [19], [20], our methodology creates robustness versus minor motion artifacts caused by moving patients or noise.

One limitation of this study despite theoretical success is related to a limited number of patients. Infrared thermography in such a setup is not easy to obtain, and a bigger cohort of patients or another dataset with similar imaging boosts the statistical significance of this study by increasing the likelihood of independent validation of the system (in place of cross-validation). A number of thermomics also can be considered as another limitation of this study. Having more thermomics helps to test the strength of the dimensionality reduction method leading to capture more reliable thermal characteristics than the current amount.

The proposed embedding system offers some improvements that can be pointed out. First, embedded Convex-NMF not only performed a factorization analysis of thermal images but also entirely eliminates the manual selection of the important bases. This substantially aids the thermographic systems and is often stated to be an issue in thermography. Second, this approach uses thermomics to obtain heterogeneous thermal patterns. Third, the suggested system alleviates the impact of motion artifacts which can be a substantial help in infrared thermography applications.

V. CONCLUSION

This study tackled one of the major challenges in the LD representation of thermal sequences, which is choosing the predominant representative basis and proposing a bell-curve embedding approach. Convex factorization analysis was examined and tested for 208 thermal breast cancer screening

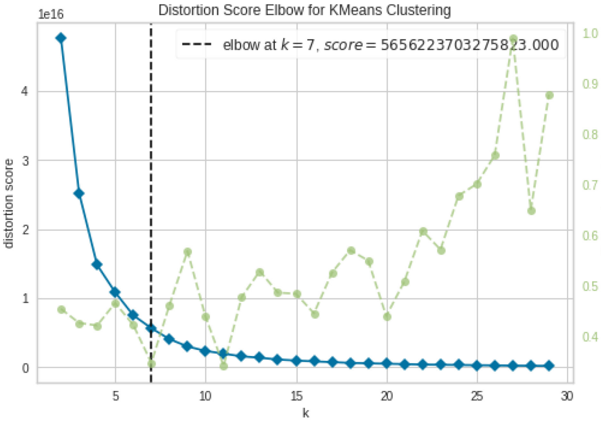

FIGURE 7. **Elbow technique to calculate the distortion score and find out the optimum number of the cluster for conventional radiomics. $k = 7$ is optimum number, shown by the graph, where we select $k = 7$ for this study.**

cases. 354 thermomics were extracted to encode thermal patterns and use them for the automatic diagnostic model. The dimensionality was lowered by applying the spectral embedding approach. The correctness of this approach was comparatively evaluated with respect to state-of-the-art thermographic methods, such as PCT, CCIPCT, NMF, Sparse PCT, Sparse NMF, and Deep semi-NMF and with Gaussian embedding. The results indicated that Gaussian and bell-curve-embedded Convex-NMF have significant functioning in maintaining thermal heterogeneity, which led to discriminating between abnormal and healthy participants yielded the accuracies of 77.6% (43.2% and 87.4%), and 76.9% (73.7% and 86.1%), respectively. The highest accuracy belonged to 77.9% (45.7% and 87.4%) with a kappa coefficient of 76.8 (43.7 and 87.6).

Future works will include more methodological development to enhance the ability to measure thermal dimensionality reduction with respect to the best predominant representation of the thermographic images. Moreover, a bigger cohort of patients would be helpful to provide the opportunity to independently validate this system and can further confirm the generalizability and reliability of this approach.

APPENDIX

In this study, we determined the optimum number of the convex factorization analysis can be obtained through the elbow approach as it is shown in Fig. 7. This approach follows the known theorem of similarity of Convex NMF to Kmeans clustering (Theorem A) [15].

Theorem A: Convex NMF is the relaxation of K means clustering.

REFERENCES

[1] R. L. Siegel, K. D. Miller, H. E. Fuchs, and A. Jemal, "Cancer statistics," *Cancer J. Clin.*, vol. 71, no. 1, pp. 7–33, 2021.

[2] B. Yousefi, H. Akbari, and X. P. V. Maldague, "Detecting vasodilation as potential diagnostic biomarker in breast cancer using deep learning-driven thermomics," *Biosensors*, vol. 10, no. 11, p. 164, 2020, doi: 10.3390/bios10110164.

[3] P. Gamagami, "Indirect signs of breast cancer: Angiogenesis study," in *Atlas of Mammography*. Cambridge, MA, USA: Blackwell Sci., 1996, pp. 321–326.

[4] T. Yahara, T. Koga, S. Yoshida, S. Nakagawa, H. Deguchi, and K. Shirouzu, "Relationship between microvessel density and thermographic hot areas in breast cancer," *Surg. Today*, vol. 33, no. 4, pp. 243–248, 2003.

[5] N. Rajic, "Principal component thermography for flaw contrast enhancement and flaw depth characterisation in composite structures," *Composite Struct.*, vol. 58, no. 4, pp. 521–528, 2022.

[6] R. Usamentiaga, Y. Mokhtari, C. Ibarra-Castanedo, M. Klein, M. Genest, and X. Maldague, "Automated dynamic inspection using active infrared thermography," *IEEE Trans. Ind. Informat.*, vol. 14, no. 12, pp. 5648–5657, Dec. 2018.

[7] S. Marinetti, L. Finesso, and E. Marsilio, "Matrix factorization methods: Application to thermal NDT/E," *NDT E Int.*, vol. 39, no. 8, pp. 611–616, 2006.

[8] K. E. Cramer and W. P. Winfree, "Fixed Eigenvector analysis of thermographic NDE data," in *Proc. Thermosense Thermal Infrared Appl. XXXIII*, 2011, pp. 225–235. [Online]. Available: https://doi.org/10.1117/12.882359

[9] B. Yousefi *et al.*, "Incremental low rank noise reduction for robust infrared tracking of body temperature during medical imaging," *Electronics*, vol. 8, no. 11, p. 1301, 2019.

[10] J. Ahmed, B. Gao, and W. L. Woo, "Wavelet-integrated alternating sparse dictionary matrix decomposition in thermal imaging CFRP defect detection," *IEEE Trans. Ind. Informat.*, vol. 15, no. 7, pp. 4033–4043, Jul. 2019.

[11] M. Barry *et al.*, "Multimodal radiothermomic biomarkers for breast cancer screening," in *Proc. Thermosense Thermal Infrared Appl. XLIV*, 2022, pp. 115–126.

[12] B. Yousefi, S. Sfarra, C. I. Castanedo, and X. P. V. Maldague, "Comparative analysis on thermal non-destructive testing imagery applying candid covariance-free incremental principal component thermography (CCIPCT)," *Infrared Phys. Technol.*, vol. 85, pp. 163–169, Sep. 2017.

[13] B. Yousefi, S. Sfarra, F. Sarasini, C. I. Castanedo, and X. P. V. Maldague, "Low-rank sparse principal component thermography (sparse-PCT): Comparative assessment on detection of subsurface defects," *Infrared Phys. Technol.*, vol. 98, pp. 278–284, May 2019.

[14] J.-Y. Wu, S. Sfarra, and Y. Yao, "Sparse principal component thermography for subsurface defect detection in composite products," *IEEE Trans. Ind. Informat.*, vol. 14, no. 12, pp. 5594–5600, Dec. 2018.

[15] C. H. Q. Ding, T. Li, and M. I. Jordan, "Convex and semi-nonnegative matrix factorizations," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 1, pp. 45–55, Jan. 2010.

[16] B. Yousefi, S. Sfarra, C. Ibarra-Castanedo, N. P. Avdelidis, and X. P. V. Maldague, "Thermography data fusion and nonnegative matrix factorization for the evaluation of cultural heritage objects and buildings," *J. Thermal Anal. Calorimetry*, vol. 136, no. 2, pp. 943–955, 2019.

[17] B. Yousefi, C. Ibarra-Castanedo, and X. P. V. Maldague, "Infrared non-destructive testing via semi-nonnegative matrix factorization," *Proceedings*, vol. 27, no. 1, p. 13, 2019.

[18] B. Yousefi, C. I. Castanedo, and X. P. V. Maldague, "Measuring heterogeneous thermal patterns in infrared-based diagnostic systems using sparse low-rank matrix approximation: Comparative study," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–9, 2021, doi: 10.1109/TIM.2020.3031129.

[19] B. Yousefi, C. I. Castanedo, and X. V. P. Maldague, "Low-rank convex/sparse thermal matrix approximation for infrared-based diagnostic system," 2020, *arXiv:2010.06784*.

[20] B. Yousefi, H. M. Sharifipour, and X. P. V. Maldague, "A diagnostic biomarker for breast cancer screening via Hilbert embedded deep low-rank matrix approximation," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–9, 2021, doi: 10.1109/TIM.2021.3085956.

[21] B. Yousefi, M. Hershman, H. C. Fernandes, and X. P. V. Maldague, "Concentrated thermomics for early diagnosis of breast cancer," *Eng. Proc.*, vol. 8, no. 1, p. 30, 2021.

[22] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[23] J. J. M. Van Griethuysen *et al.*, "Computational radiomics system to decode the radiographic phenotype," *Cancer Res.*, vol. 77, no. 21, pp. e104–e107, 2017.

[24] L. F. Silva *et al.*, "A new database for breast research with infrared image," *J. Med. Imag. Health Inform.*, vol. 4, no. 1, pp. 92–100, 2014.

[25] V. Berisha *et al.*, "Digital medicine and the curse of dimensionality," *NPJ Digit. Med.*, vol. 4, no. 1, pp. 1–8, 2021.

[26] C. Parmar, J. D. Barry, A. Hosny, J. Quackenbush, and H. J. W. L. Aerts, "Data analysis strategies in medical imaging," *Clin. Cancer Res.*, vol. 24, no. 15, pp. 3492–3499, 2018.