

Solution Brief

Medical Imaging, AI-Based Breast Cancer Screening

INTEL LOGO
insert official asset

Intel® Core™ Ultra Processors Power Niramai's Thermalytix® Breast Cancer Screening at the Edge

Niramai's Thermalytix screening solution — migrated from the cloud to Intel® Core™ Ultra edge platforms — delivers the identical AI risk assessment on premises, with faster report turnaround and no dependence on internet connectivity.

Prepared jointly by Niramai Health Analytix Pvt Ltd, India and Intel Corporation.

Executive Summary

Niramai Health Analytix, a Bengaluru-based deep-tech healthcare startup, developed a novel, AI-based diagnostic solution called Thermalytix® that enables radiation-free, non-invasive, privacy-conscious, and affordable breast screening, suitable for women of all age groups and breast densities. Since its inception, Thermalytix has been deployed as a cloud-native platform, where thermal images acquired at screening centres are securely uploaded to a central cloud infrastructure for AI inference, report generation, storage, and clinical review. This architecture enabled rapid deployment, centralised model updates, scalable computing resources, and seamless integration with multiple screening locations. However, reliance on cloud connectivity presents practical challenges in many real-world screening environments. Rural and remote clinics often experience unreliable or low-bandwidth internet access, community screening camps may have intermittent connectivity, and several hospitals and government institutions enforce strict data-residency and cybersecurity policies that restrict patient data from leaving their local infrastructure. These constraints can increase report turnaround times, interrupt screening workflows, and limit the deployment of AI-assisted screening in settings where it is needed most.

To overcome these challenges, Niramai collaborated with Intel to develop an edge deployment of the complete, clinically validated Thermalytix software stack from the cloud to Intel® Core™ Ultra edge platforms. The migration preserved the entire micro-services based architecture, enabling all application services to execute locally on the edge without requiring continuous cloud connectivity. By maintaining the same validated software stack and service interactions, the edge deployment produced identical clinical risk assessments to the cloud implementation, a critical requirement for a medical screening product. In benchmark testing, the end-to-end screening API pipeline completed approximately 4× faster on an Intel Core Ultra X7-based edge system (and approximately 3.5× faster on Intel Core Ultra 7) than on the reference cloud deployment, with all patient data remaining on premises.¹ A prototype of this edge AI solution for breast cancer screening was showcased at the Intel Client Ecosystem Symposium and Edge Solution Summit, Taipei, May 2026.

This brief describes the motivation for the migration, the edge architecture, the engineering work involved, the validation methodology, and measured performance across Intel platforms — along with lessons learned and the roadmap ahead.

Expanding Access to Breast Cancer Screening

Breast cancer is the most common cancer among women worldwide and claims more than 680,000 lives each year [1]. In the United States, an estimated 321,910 women will be diagnosed with invasive breast cancer and 42,140 women will die from the disease in 2026. In India, over 220,000 new breast cancer cases and 80,000 deaths are reported

annually. Survival depends heavily on early detection, yet millions of women lack access to timely breast health screening.

Current screening pathways face significant barriers, particularly in low- and middle-income countries. Population-based mammography programs require expensive infrastructure, trained personnel, and regular access to healthcare facilities, limiting their reach in many communities. As a result, clinical breast examination remains the primary screening approach in several settings, despite its limited sensitivity for early disease.

An important unmet need also exists for younger women and women with dense breasts. Younger women are generally not included in routine population screening programs, while dense breast tissue can reduce the effectiveness of conventional imaging approaches. These populations often experience delayed detection despite having a substantial risk of developing breast cancer. This challenge extends beyond low- and middle-income countries, driving global interest in innovative technologies that complement existing screening pathways and improve access to early breast health assessment.

There is an urgent need for a scalable, affordable, and accessible breast health screening solution that can expand access to early risk assessment across diverse populations and enable timely referral for diagnostic evaluation.

Thermalytix®: AI-Powered Breast Health Screening

Thermalytix® is Niramai Health Analytix's patented AI-powered breast health screening platform that integrates a portable high-resolution infrared thermal imaging device with a proprietary AI analytics engine. The platform automatically analyses physiological thermal biomarkers including thermal asymmetry, vascular patterns, and focal temperature variations to generate an objective, quantitative breast-health risk assessment that supports clinical decision-making.

Unlike conventional imaging modalities, Thermalytix is completely non-contact, non-invasive, and radiation-free, requiring no breast compression during image acquisition [2]. A screening examination can typically be completed within a few minutes using a portable thermal camera, enabling deployment across tertiary hospitals, diagnostic centres, primary healthcare facilities, and community outreach programmes. Its portable design and simple acquisition workflow allow trained healthcare workers to perform breast health screening at the point of care, extending access beyond specialised imaging centres.

Over the past decade, Thermalytix has evolved into a mature AI platform through continuous algorithm development, large-scale clinical validation, and real-world deployment. The technology is protected by a strong intellectual property portfolio comprising 40 granted patents spanning thermal imaging, artificial intelligence, image analytics, and automated breast-health assessment. The platform has also been

described in more than 50 peer-reviewed scientific publications and has received multiple regulatory clearances, demonstrating both scientific credibility and commercial readiness.

Thermalytix has been evaluated through multiple prospective and retrospective clinical studies conducted across diverse healthcare settings involving thousands of women. Published studies have consistently reported clinical sensitivities exceeding 95%, specificities exceeding 80%, and a negative predictive value greater than 98%, supporting its role as a frontline breast health screening and triage solution [3-9]. Because Thermalytix analyses physiological thermal biomarkers rather than anatomical tissue density, it has demonstrated robust performance in younger women and women with dense breasts—populations for whom conventional mammography presents recognised screening challenges.

Today, Thermalytix has been deployed across 300+ healthcare facilities in multiple countries and has supported breast health screening for more than 550,000 women. Its clinically validated AI models, mature microservices-based software architecture, and proven deployment at scale provided a strong foundation for extending the platform from a cloud-native deployment to Intel® Core™ Ultra edge systems while preserving identical clinical functionality.

As deployment expanded into increasingly diverse screening environments, the need for greater local autonomy, reduced dependence on connectivity, and support for stricter data-governance requirements became increasingly important. These requirements motivated the extension of Thermalytix from its established cloud-native architecture to an edge-based deployment model. This document details the design and development of this edge-based Thermalytix solution and its deployment on Intel® Core™ Ultra systems, while preserving the established clinical workflow and functionality.

Initial Cloud-Native Deployment of Thermalytix

Thermalytix was originally developed as a cloud-native, microservices-based platform, with AI processing and application services hosted in the cloud. During a screening examination, thermal images acquired at the point of care are securely uploaded to cloud-hosted microservices that perform image preprocessing, breast segmentation, AI inference, report generation, user authentication, and data management. The generated breast-health report is then returned to the screening workstation for clinical review.

The cloud-native architecture enabled rapid deployment across multiple screening locations while simplifying software maintenance, centralised management, and AI model updates. By hosting application services centrally, all deployments could operate using the latest validated software stack without requiring local software installation or maintenance. The architecture also provided elastic compute resources to support growing screening volumes and simplified operational monitoring across geographically distributed deployments.

This deployment model successfully supported large-scale breast health screening across hospitals, diagnostic centres, and community screening programmes. As the platform expanded into increasingly diverse healthcare environments, however, its dependence on cloud connectivity introduced several operational constraints. Thermal images had to be uploaded for processing and reports downloaded before the patient could complete the screening workflow. In community and rural settings with intermittent internet connectivity, this introduced an operational dependency that could interrupt screening. Cloud availability also created a shared dependency across deployments, while data transfer added latency and cloud infrastructure incurred recurring costs at each site. These requirements motivated the development of an edge-based deployment.

Motivation for Edge Deployment

While the cloud-native architecture continues to provide an efficient deployment model for many healthcare providers, certain clinical environments require an alternative deployment option that can operate independently of continuous cloud connectivity while preserving the same clinically validated screening workflow.

An edge deployment addresses these requirements by enabling the complete Thermalytix platform to execute locally within the healthcare facility, offering several operational and technical advantages:

- **On-premises data processing:** On-premises data processing: Patient images, reports, and associated clinical data are processed and stored within the healthcare facility, helping support local data-residency requirements, as well as institutional privacy, cybersecurity, and data-governance policies. Patient data does not need to be transferred to a cloud environment for routine Thermalytix processing.
- **Offline operation:** Screening can continue without continuous internet connectivity, enabling reliable deployment in rural clinics, community outreach programmes, mobile screening units, and other bandwidth-constrained environments.
- **Lower and more predictable latency:** Local AI inference eliminates the need to transfer images to the cloud for analysis, resulting in faster report generation and a more responsive screening workflow.
- **Operational resilience:** Independent edge deployments continue to function even during internet or cloud service interruptions, ensuring uninterrupted screening operations.
- **Deployment flexibility:** Healthcare providers can choose the deployment model—cloud, edge, or hybrid—that best aligns with their infrastructure, connectivity, operational, and regulatory requirements.



Figure 1. Deployment setup — thermal camera and Intel Core Ultra edge box in a screening-room configuration.

Thermalytix on Intel Edge Platforms

To address the operational requirements of healthcare providers seeking local AI execution, Niramai collaborated with Intel to introduce an edge deployment option for Thermalytix on Intel® Core™ Ultra platforms. The objective was to enable the complete, clinically validated Thermalytix platform to operate entirely within a healthcare facility while preserving the same screening workflow, AI models, and clinical outputs as the cloud deployment. Figure 1 shows the edge deployment setup, in which the Intel edge computing box and thermal camera are mounted on a tripod to form a unified, point-of-care screening device.

The edge deployment retains the existing microservices-based software architecture, allowing image acquisition, preprocessing, breast segmentation, AI inference, report generation, authentication, and data management services to execute locally on the Intel® Core™ Ultra system (Figure 2). During a screening examination, thermal images acquired at the point of care are transferred over the local network to the edge platform, where the complete AI pipeline executes without requiring cloud connectivity. The generated breast-health report is returned immediately to the screening workstation for clinical review, while all patient images, intermediate processing artefacts, and reports remain within the healthcare facility.

In the initial deployment topology, the thermal camera continues to connect to the technician laptop as in the cloud-based workflow. The laptop accesses the edge platform through a URL on the local network, with both Wi-Fi and LAN configurations tested, including combined configurations. Upload, analysis, and storage are performed locally on the edge box, with no patient data leaving the local network. The edge

box is configured with two Windows user accounts: an administrator account for installation and maintenance, and a separate user account that allows the box to also function as the screening workstation when required.

Implementing Thermalytix on Intel® Core™ Ultra platforms required more than simply deploying the existing application onto new hardware. While the clinically validated AI models and core microservices remained unchanged, the surrounding application infrastructure had to be adapted to support secure, reliable, and fully local operation within a healthcare facility. The engineering effort therefore focused on replacing cloud-dependent infrastructure with local services while maintaining identical application behaviour and clinical outputs.

network security. Consequently, network integration became a site-specific deployment activity rather than a single configuration that could be applied uniformly across all installations.

The deployment was also designed to support a range of hardware configurations, including systems with 8 GB, 32 GB, and 64 GB of RAM, rather than being tied to a single reference configuration. Because this represented the first deployment of the validated cloud-based Thermalytix pipeline onto edge hardware, each service and its interactions had to be re-validated in the local environment. This validation ensured that the transition to edge execution did not alter the established clinical workflow, AI outputs, or report generation.

In the initial edge deployment, AI inference runs entirely on the CPU. No GPU, NPU, or OpenVINO-based optimization is currently used,

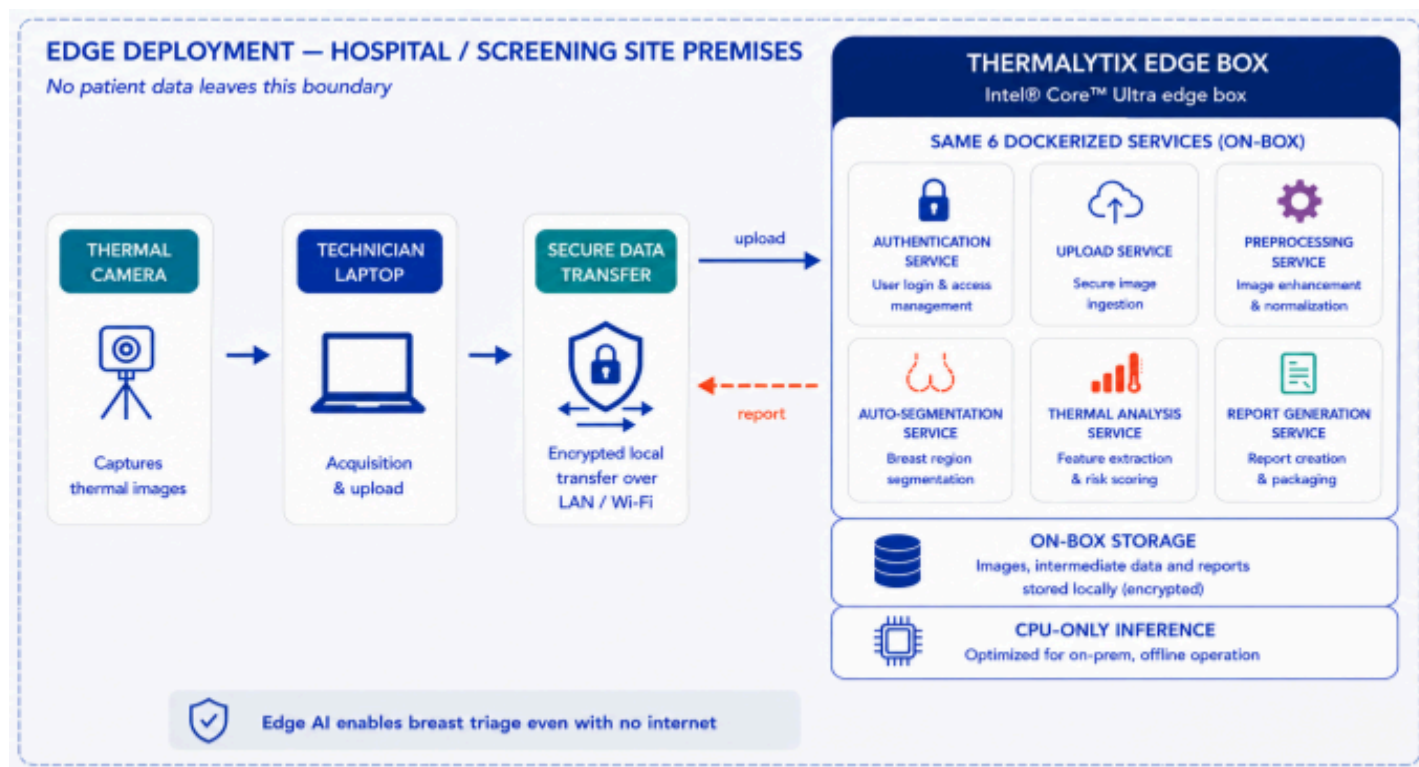


Figure 2. Edge architecture — thermal camera → technician laptop → edge box (local APIs, storage) over Wi-Fi/LAN.

Key engineering activities included:

- **Local networking** — The technician laptop that receives images from the thermal camera accesses the edge box through a URL on the same local network rather than a public cloud endpoint. Both Wi-Fi and LAN connectivity were tested independently as well as in combined configurations.
- **Local data management** — Uploaded thermal images, intermediate processing artefacts, and generated reports are stored entirely on the edge box, with no cloud fallback or external data transfer during the screening workflow.
- **Headless device access** — The NUC-class edge box is typically deployed without a dedicated monitor or keyboard. Remote administration therefore had to be planned and configured as part of the deployment rather than treated as an after-installation requirement.
- **Account separation** — Two Windows user accounts are configured on each edge system: an administrator account for installation and maintenance, and a separate user account that allows the device to function as a screening workstation when required.

Hospital environments introduced additional deployment considerations beyond the software architecture itself. Hospital Wi-Fi networks are typically restricted by institutional IT policies, and integrating an edge system into these environments required the device to be discoverable and reachable by the technician laptop without compromising existing

despite the presence of integrated GPU and NPU capabilities on the evaluated Intel® Core™ Ultra platforms. This was a deliberate engineering choice: keeping the existing clinically validated AI pipeline unchanged enabled a direct comparison between cloud and edge deployments without introducing changes to the inference stack as an additional variable. During testing, observed CPU utilization remained low and the system achieved the required performance without platform-specific acceleration. OpenVINO and NPU-based optimization therefore remain potential avenues for further improving efficiency, particularly for lower-power edge configurations.

By preserving the same validated software stack and AI models across cloud and edge deployments, healthcare providers can select the deployment model that best meets their connectivity, operational, and data-governance requirements without changing the clinical workflow or screening experience..

Platform	Processor	RAM	Notes
Panther Lake	Intel® Core™ Ultra X7 358H (3.70 GHz), Intel Arc B390 iGPU	64 GB	Windows, x64
Lunar Lake	Intel® Core™ Ultra 7 258V (2.20 GHz), Intel Arc 140V iGPU	32 GB	Windows, x64

Validating Clinical Equivalence

For a medical screening product, the question that mattered most in this migration was not how fast the edge box ran — it was whether it produced the same clinical output as the cloud system it was replacing. Because the identical Dockerized software stack and the same AI models run on both cloud and edge, with no model changes made during migration, the Thermalytix risk assessment output is expected to be identical regardless of where it runs. Validating that expectation, rather than chasing speed, was the primary success criterion for this phase of the project.

Service	Endpoint	Panther Lake (edge, ms)	Lunar Lake (edge, ms)	Cloud (ms)
Authentication	/api/login	69	75	144
Image upload	/api/upload	80	92	539
Image conversion	/api/preprocess	277	408	1,953
Auto-segmentation	/api/autosegment	586	689	2,158
Analysis	/api/report	4,950	5,695	19,465
Fetch report	/api/uploads	25	21	385
Total (6 APIs)		≈6.0 s	≈7.0 s	≈24.6 s

The validation method was direct: the same input thermal images for two test subjects were run through three separate deployments — the existing cloud system, the Panther Lake edge configuration, and the Lunar Lake edge configuration — and the resulting Thermalytix risk scores were compared across all three.

Result: the predicted risk scores matched exactly across all deployments — a score of 2 for one subject and a score of 5 for the other, identical on cloud, Panther Lake, and Lunar Lake. This is the central claim of the migration: the same validated pipeline, moved to different hardware, produces the same screening result.

This two-subject test served as a benchmark comparison, not a full clinical regression suite; broader validation scope and methodology are provided by Niramai QA as noted above. What it establishes is the foundation the rest of this paper builds on: because equivalence was preserved, the performance gains discussed next are a genuine bonus on top of a system that behaves identically to the one already validated in the cloud, not a trade-off against clinical reliability.



Figure 3. Risk scores by subject across cloud, Panther Lake, and Lunar Lake deployments.

Performance Evaluation

With equivalence established, the team measured per-API latency across the same two test subjects, comparing Panther Lake and Lunar Lake edge deployments against the cloud reference. All values are average latency in milliseconds across the two test subjects.

These figures should be read with one caveat in mind: the cloud reference used in this comparison was a cost-optimized configuration,

not a performance-tuned one. The results are indicative of what local processing removes from the pipeline — network round trips, upload and download time — rather than a like-for-like hardware benchmark, and performance varies by configuration.

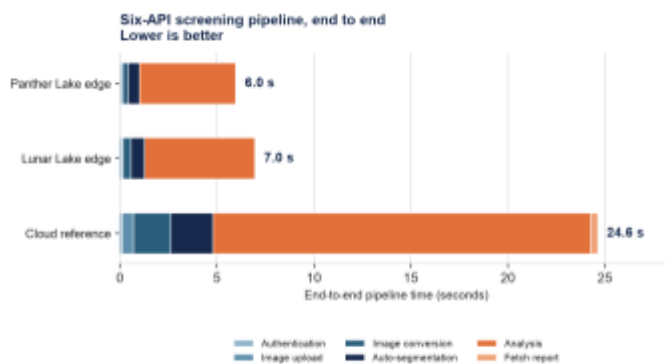


Figure 4. End-to-end pipeline time by deployment, broken down by service.

In practical terms, the numbers translate into a full screening report returned in seconds, over the local hospital network, with no dependency on internet quality. For a screening workflow that has historically been slowed by uploads and downloads over patchy or shared connectivity, that is the change technicians are most likely to notice day to day.

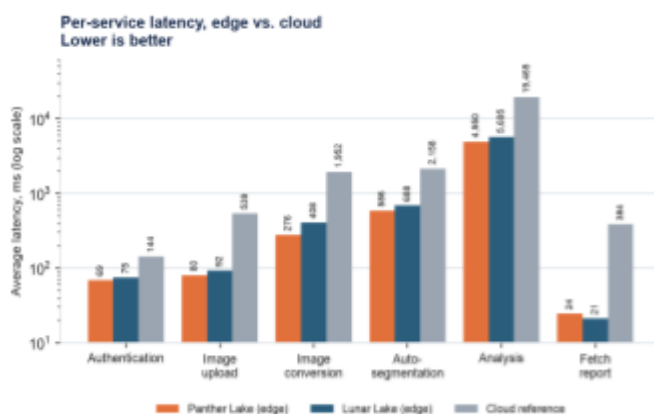


Figure 5. Latency comparison chart — per-API latency, Panther Lake vs. Lunar Lake vs. cloud.

Deployment Lessons

- **Preserving the validated pipeline enabled a direct equivalence assessment.** Keeping the AI models and inference logic unchanged allowed the team to evaluate the edge implementation without introducing changes to the clinical AI pipeline..

- **Local deployment simplifies data governance, while network integration remains an important consideration.** Keeping patient data entirely on the edge system addresses on-premises data requirements, while integration with restricted hospital networks requires careful local network configuration.
- **Edge deployment introduces new operational considerations at scale.** Headless device administration, software updates, monitoring, backup, and fleet management become increasingly important as independent edge systems are deployed across multiple healthcare facilities.

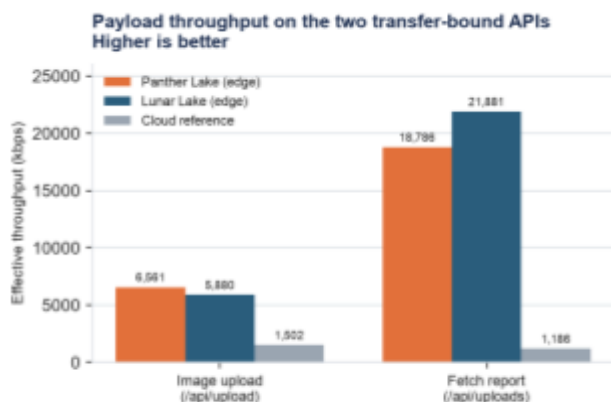


Figure 6. Throughput comparison chart — effective throughput by API, edge vs. cloud.

What's Next

The initial implementation validated the edge approach on Panther Lake and Lunar Lake platforms, which provide substantial compute capacity for the existing Thermalytix workload. Evaluation is now extending to lower-power and lower-cost Intel® platforms to determine how the same validated pipeline performs on more constrained configurations that may be suitable for high-volume and resource-limited screening environments.

In parallel, the team is exploring OpenVINO™ and NPU-based optimization for these lower-power platforms. While such optimization was not necessary for the initial implementations given the available CPU capacity, hardware acceleration may become increasingly valuable as the target platforms become smaller and more power-efficient.

As the number of edge deployments grows, the next stage will extend beyond the individual device to the broader edge ecosystem, including fleet management, remote software updates, centralized monitoring, data backup, and deployment models suited to kiosk-based and community screening environments. These capabilities can help extend the benefits of local AI beyond individual clinical workstations and support scalable deployment of Thermalytix across diverse healthcare settings.

This version keeps the real lessons and limitations—which actually makes the document more credible—but frames them as engineering and scalability considerations rather than weaknesses.

Conclusion

The deployment of Thermalytix on Intel® Core™ Ultra edge platforms demonstrates that clinically validated medical AI can be brought directly to the point of care without changing its established AI models, clinical workflow, or outputs. By retaining the validated Dockerized software stack while enabling local execution, Niramai achieved a fully on-premises deployment that reduces dependence on connectivity and keeps patient data within the healthcare facility. The complete Thermalytix workload was successfully executed using CPU-only inference, demonstrating the capability of Intel® Core™ Ultra processors to support the existing clinical pipeline while leaving integrated GPU, NPU, and OpenVINO™ capabilities available for future optimization. This provides a practical foundation for bringing trusted medical AI to more healthcare environments where privacy, resilience, and local processing are essential.

References

- [1] Global Cancer Observatory (GCO). International Agency for Research on Cancer (IARC), World Health Organization (WHO). Available at: <https://gco.iarc.who.int/en>. Accessed August 25, 2026.
- [2] Kakileti ST, Madhu H, Subramoni T, Manjunath G. Thermalytix: using AI to save lives. XRDS: Crossroads, The ACM Magazine for Students. 2020 Apr 2;26(3):38-41.
- [3] Kakileti ST, Madhu HJ, Krishnan L, Manjunath G, Sampangi S, Ramprakash HV. Observational study to evaluate the clinical efficacy of thermalytix for detecting breast cancer in symptomatic and asymptomatic women. JCO global oncology. 2020 Oct;6:1472-80.
- [4] Bansal R, Collison S, Krishnan L, Aggarwal B, Vidyasagar M, Kakileti ST, Manjunath G. A prospective evaluation of breast thermography enhanced by a novel machine learning technique for screening breast abnormalities in a general population of women presenting to a secondary care hospital. Frontiers in Artificial Intelligence. 2023 Jan 4;5:1050803.
- [5] Singh, A., Bhat, V., Sudhakar, S., Namachivayam, A., Gangadharan, C., Pulchan, C. and Sigamani, A., 2021. Multicentric study to evaluate the effectiveness of Thermalytix as compared with standard screening modalities in subjects who show possible symptoms of suspected breast cancer. BMJ open, 11(10), p.e052098.
- [6] Adapa K, Gupta A, Singh S, Kaur H, Trikha A, Sharma A, Rahul K. A real world evaluation of an innovative artificial intelligence tool for population-level breast cancer screening. NPJ digital medicine. 2025 Jan 2;8(1):2.
- [7] Kakileti ST, Madhu HJ, Manjunath G, Wee L, Dekker A, Sampangi S. Personalized risk prediction for breast cancer pre-screening using artificial intelligence and thermal radiomics. Artificial intelligence in medicine. 2020 May 1;105:101854.
- [8] Kakileti ST, Sampangi S, Ramprakash HV, Manjunath G. Evaluating Clinical Efficacy of Thermalytix as a Breast Cancer Screening Tool in Young Women. Journal of Young Womens Breast Cancer and Health. 2024 Jan 1;1(1&2):58-65.
- [9] Ramprasad J, Krishnan L, Gangadharan C, Deshpande G, Madhu H, Kakileti ST, Manjunath G. Performance of artificial intelligence-based breast cancer screening in a community setting: a real-world evaluation study. The Lancet Oncology. 2022 Jul 1;23:S20.