WESTERN caspian UNIVERSITY

e-ISSN 3105-4293
p-ISSN 3105-4285

Journal of Nanotechnology and Innovative Materials

Vol 2, № 1, 2026

Baku - 2026

Western Caspian University

e-ISSN 3105-4293

p-ISSN 3105-4285

Journal of Nanotechnology and Innovative Materials

Vol 2, No 1, 2026

DOI:10.54414/VWTT8467

https://jnim.wcu.edu.az/

CONTENTS

https://doi.org/10.54414/ERVK9786

Journal of Nanotechnology and Innovative Materials

Research Article

Explainable Vision Transformer-Based Automated Malaria Detection from Blood Smear Images: Toward Intelligent Nanotechnology-Enabled Diagnostic Systems

Sabina M. Nobari[1], Sabina R. Azizova[1], and Shokrollah A. Ghadyani[2]

[1]Department of Information Technologies, School of Advanced Technologies and Innovation Engineering, Western Caspian University, Ahmad Rajabli 17A, III Parallel, AZ1001 Baku, Azerbaijan
[2]Tekdata Co., No. 12, Marzbannameh Alley, Mofatteh Street, Tehran, Iran

Received: 23.06.2026 Accepted: 28.06.2026 Published: 31.07.2026

https://doi.org/10.54414/LCGJ5320

Abstract

Malaria remains one of the world's most significant parasitic diseases, requiring rapid and accurate diagnosis to reduce mortality. Recent advances in digital microscopy, artificial intelligence, and nanotechnology-enabled diagnostic platforms have created new opportunities for automated malaria detection. In particular, nano-enabled biosensors and lab-on-chip technologies increasingly rely on intelligent image analysis algorithms to improve diagnostic performance. In this study, convolutional neural network architectures (ResNet50, EfficientNetB0, InceptionResNetV2, and Xception) and Vision Transformer models (ViT-Base, Swin Transformer, DeiT, and Pyramid Vision Transformer) were evaluated for multiclass malaria detection using microscopic blood smear images. A five-fold cross-validation strategy was employed, and model interpretability was enhanced using Gradient-weighted Class Activation Mapping (Grad-CAM). Among the CNN models, Xception achieved an accuracy of 98.10%, while the Swin Transformer demonstrated comparable performance among transformer-based architectures. The results indicate that combining explainable artificial intelligence with transformer-based models improves both classification accuracy and model transparency, supporting their future integration into nano-enabled diagnostic systems, intelligent digital microscopy, and point-of-care biomedical devices.

Keywords: malaria diagnosis, deep learning, vision transformer, explainable artificial intelligence, nanotechnology, nanobiosensors, medical image analysis, digital microscopy

1. Introduction

Malaria remains one of the most significant parasitic infectious diseases worldwide and continues to pose a major public health challenge, particularly in tropical and subtropical regions. Millions of people get this disease every year. According to the World Health Organization (WHO, 2024), in 2023 there were 263 million malaria cases worldwide and about 597 thousand deaths worldwide. The main cause of the disease is various types of Plasmodium parasites (most commonly P. falciparum and P. vivax). Malaria is transmitted to humans through the bite of infected female *Anopheles* mosquitoes [1]. Effective treatment of malaria in full and on time and the diagnosis of the infection are of absolute importance. Since the beginning of the traditional microscopic diagnostic method of the 20th century, the "gold standard" has been established; this method is open to allowing morphological changes in Plasmodium infection as a result of Erythrocytes (red blood cells-RBC). In this method, stained blood is checked under a microscope; then, whichever parasite number in erythrocytes is considered high accuracy. This method is a process that is labor-intensive and operator-dependent process, although it requires a significant amount of time. Sources say that human errors in clinics can affect the results, and serious effects can be demonstrated [1], [2].

Molecular methods such as polymerase chain reaction (PCR) and rapid diagnostic tests (RDT) also improve, but these methods require high-level laboratory equipment and are expensive, so both areas need large-scale and appropriate-level implementation. Therefore, increases in systems that require rapid results, automated, accurate, and cost-effective malaria detection [3], [4]. In recent years, Computer Vision (CV) and Deep Learning (DL) have significantly advanced the automated analysis of microscopic blood smear images. Various deep learning architectures, including Convolutional Neural Networks (CNNs), Faster R-CNN, YOLO, and Mask R-CNN, have been successfully applied to automated malaria detection [5]. These models automatically extract discriminative image features from microscopic blood smear images, enabling accurate classification of malaria parasites. Many CNN-based research systems have reported that malaria detection is between 95% and 99% accurate [6], [7]. However, CNN models have limitations in capturing long-range dependencies and understanding global contextual relationships between distant pixels and image regions. These limitations restrict their ability to effectively model complex spatial relationships within images.

To overcome these limitations, a new generation of models inspired by Transformer architectures has been introduced. Originally developed for natural language processing (NLP), these models have demonstrated remarkable performance in computer vision tasks. These models show an image of various parts between long-distance dependencies and global attention, the open relationships [8].

The promising results of the images in the analysis, including the architecture of Vision Transformer (ViT), Data Efficient Image Transformer (DeiT), Pyramid Vision Transformer (PVT), and Swin Transformer (malaria parasites in the testpit) [9], [10].

One of the major limitations of AI in medical diagnosis is the lack of interpretability (or explainability). Clinical specialists need not only a complete, but also a visualized explanation of the decision-making process. To this end, Gradient is mainly used in Class Activation Mapping (Grad-CAM) methods, but also helps decide on the most impactful areas highlighted by the model. This is both an increase in the trust of doctors in AI systems, as well as facilitating error analysis and improvement of algorithms [11].

Moreover, in recent years, nanotechnology has significantly contributed to the development of rapid malaria diagnostic systems through nanoparticle-based biosensors, plasmonic sensing platforms, microfluidic lab-on-chip devices, and portable nano-enabled imaging technologies [12]. These emerging diagnostic tools generate high-resolution microscopic data that require intelligent computational algorithms for accurate interpretation. Consequently, artificial intelligence and deep learning have become essential complementary technologies for improving the diagnostic performance of nanotechnology-based medical devices.

1.1. Research Gap

In the last decade, a large number of CNN-based research conducted for the discovery of automated malaria detection is also directed, the majority of these studies have focused only on binary classification of infected and uninfected cells, and the examination of multiple analytic time patterns that were revealed to be far deeper than analytic methods that could be construed.

CNN models perform very well in situations where global additions and ink tissue are fully exposed to learning challenges. This leads to a reduction in accuracy, especially when the images are noisy or there are lighting differences. This factor is also limited in the number yet applied by the researchers. The development of explainable Artificial Intelligence (Explainable AI, XAI) approaches, by making the decision-making processes of artificial intelligence models more transparent and understandable, increases the trust of medical professionals in these systems and makes a significant contribution to the effective use of artificial intelligence in clinical applications. The objective of this study is to develop and evaluate an explainable artificial intelligence framework for automated malaria detection by comparing CNN- and Vision Transformer-based architectures on microscopic blood smear images. The article describes classic CNN architectures, advanced Transformer-based approaches, and techniques such as Grad-CAM to identify future development aspects for significant malaria diagnostics through artificial intelligence research. Although recent studies have demonstrated the effectiveness of CNN and Vision Transformer architectures for malaria detection, comparative evaluations incorporating explainable artificial intelligence remain limited.

Furthermore, few studies have discussed how these AI models can complement emerging nanotechnology-based diagnostic platforms, including nano-enabled biosensors and lab-on-chip systems. Therefore, this study evaluates multiple deep learning architectures while emphasizing their potential integration into future intelligent nanodiagnostic devices.

1.2. Background

In recent years, significant improvements in Computer Vision (CV) and Deep Learning (DL) have changed the landscape by using microscopic blood descriptions to automatically diagnose malaria. Both open-source information databases (e.g., NIH Malaria thin smear dataset) and a large number of studies, both local to the given use, were able to pinpoint the CNN-based model Plasmodium hybrid with the Convolutional Neural Networks (CNN) and infected red blood cells.

For example, the first research for classic CNN architectures and transfer learning methods achieved high accuracy in separating 95% more than infected and healthy cells by more than 95% with the ResNet50 model [13]. At the same time [14], accuracy was achieved with 97.57% detection using the EfficientNet k-fold cross-validation method. Subsequently, CNN models combine learning methods, achieving more than 98% accuracy in the resulting high-performance datasets [15], [16], [17].

Despite these successes, CNN models are still restricted in a row. They often face differences in staining due to characteristics and image information; devices sensitive to data imbalance apply to these real clinical models. Performance decreases when models are applied to new datasets (domain/field shift), and researchers have proposed two main approaches to address this issue.

1. Resources are limited, which health services around them use for light, quantity, and mobility to suit model preparation devices.
2. Image pixels between global dependencies-based architectures application for Model Transformer.

Medical visualization Vision Transformer (ViT) and its/derivatives Data-efficient Image Transformer (DeiT), Pyramid Vision Transformer (PVT), and Swin Transformer, pay close attention to the center. These models employ self-attention mechanisms to capture long-range dependencies and global contextual information, enabling more effective image feature extraction than conventional convolutional neural networks. Extensive studies have shown that Vision Transformer models achieve strong results on tasks such as classification, segmentation, and cell identification in medical imaging [18], [19]. Recent studies have further demonstrated the effectiveness of Vision Transformer architectures for 2D biomedical image classification, confirming their applicability to medical imaging tasks [20]. From this supplement, the attention mechanism more prominently combines parasite structures in the early stages of infection with CNN [7]. At the same time, multi-stage malaria detection was prepared for the hybrid CNN–ViT model, and 99% accuracy was achieved. This result is due to combining the performance of CNN and ViT, which is important to the extent that CNNs and ViT's ability to remove local properties with a global attention mechanism [21]. In addition to the other benefits of accuracy, intelligence (AI) is the main application of artificial medicine, explaining and clinical reliability. Despite the high diagnostic accuracy of deep learning models, their decision-making process is often considered a "black box," limiting their acceptance in clinical practice. Healthcare professionals require not only accurate predictions but also interpretable explanations of how these predictions are generated. To address this challenge, Explainable Artificial Intelligence (XAI) techniques, particularly Gradient-weighted Class Activation Mapping (Grad-CAM), have been widely adopted to visualize the image regions that contribute most to a model's predictions. Improved variants, such as Grad-CAM++ and DropKey Grad-CAM, further enhance localization accuracy and visualization quality. Recent studies have demonstrated that integrating Grad-CAM with both Convolutional Neural Networks (CNNs) and Vision Transformer (ViT) architectures significantly improves model interpretability and diagnostic transparency, enabling clinicians to better understand, validate, and trust AI-assisted medical diagnoses [22], [23].

1.3. Role of Nanotechnology in Malaria Diagnosis

Nanotechnology has emerged as a promising approach for improving malaria diagnosis by enabling rapid, sensitive, and portable diagnostic platforms [24]. Gold nanoparticles, magnetic nanoparticles, quantum dots, and nanostructured biosensors have been extensively investigated for detecting Plasmodium antigens, DNA, and hemozoin biomarkers. In addition, lab-on-chip and microfluidic systems integrated with nanoscale materials facilitate point-of-care diagnosis in resource-limited settings. Although these technologies significantly improve sample preparation and signal detection, automated interpretation of microscopic and biosensor-generated images remains challenging. Therefore, explainable deep learning models, such as Vision Transformers combined with Grad-CAM, provide valuable computational support for next-generation nanotechnology-enabled malaria diagnostic systems.

2. Materials and Methods

2.1. Dataset

In this study, a publicly available malaria dataset obtained from the literature [25] was used to classify different *Plasmodium* species. The dataset consists of microscopic images of peripheral blood smear samples acquired using a light microscope. The dataset consists of four different types of malaria parasites: P. falciparum, P. malariae, P. ovale, and P. vivax, and a total of 210 images. The images are 2598x1944 pixels and have a TIFF extension. Images in the dataset are shown in Figure 1.

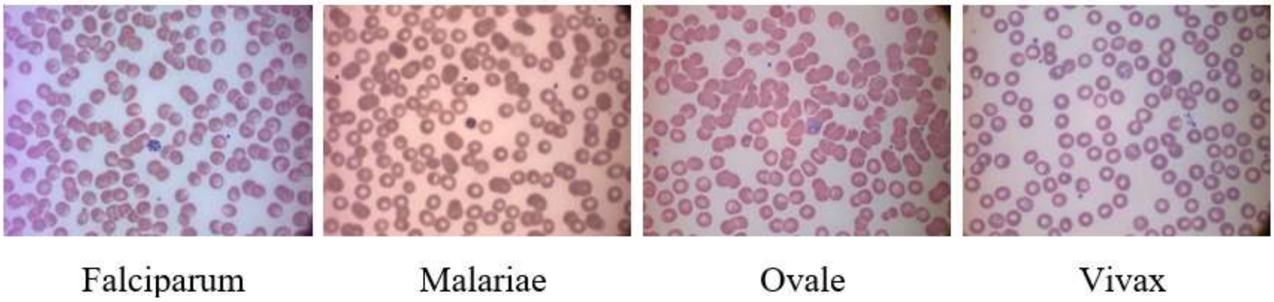

Figure 1. Sample images in the dataset.

2.2. Convolutional Neural Network

Transfer learning has been widely adopted in medical image analysis because it enables convolutional neural networks to achieve high classification accuracy even when only limited annotated datasets are available [26]. This strategy utilizes four distinct convolutional neural network architectures, ResNet50, EfficientNetB0, InceptionResNetV2, and Xception, to enhance the model's generalization capability and classification accuracy on limited medical datasets. By reusing learned feature representations, transfer learning enables models pre-trained on large-scale datasets like ImageNet to effectively adapt to domain-specific tasks, thereby reducing training time and the risk of overfitting.

ResNet50 provides redundant shortcut connections to overcome gradient problems lost in deep networks, and ensures stable training of very deep architectures [27]. EfficientNetB0 uses a compound scaling method that evenly balances network depth, width, and resolution to achieve high performance with optimal computing efficiency [28]. InceptionResNetV2 now combines Inception modules with links, offering both multi-scale feature extraction and improved convergence stability [29]. Meanwhile, Xception extends the Inception concept using deep detachable folds that improve computational efficiency and can disentangle channels [30]. Recent studies have also demonstrated that modern convolutional architectures, such as ConvNeXt, achieve competitive performance while narrowing the gap between CNNs and Vision Transformers in image classification tasks [31].

2.3. Vision Transformer

Natural language processing, created primarily from the Transformer architecture, inspired ViT, revolutionized computer vision. The evolution of Vision Transformer (ViT) from the use of layers; instead, the introduction breaks the description into small pieces (patches), each part converts to a vector, and these vectors make the global and local dependencies by means of transformer encoder modeling processing.

2.4. Self-Attention Mechanism

One of the fundamental differences between Vision Transformers (ViTs) and Convolutional Neural Networks (CNNs) is the use of the self-attention mechanism instead of convolutional filters for feature extraction. Self-attention enables ViTs to capture both local and long-range dependencies by modeling relationships among image patches across the entire image. Consequently, ViTs can effectively learn global contextual information, overcoming the limited receptive field of conventional CNNs. This capability has led to superior performance in many computer vision tasks, particularly when large training datasets are available. However, some studies have reported that CNNs may still outperform ViTs in capturing subtle local features when training data are limited or fine-grained image details are critical [32].

Attention within the ViT architecture is calculated as in Equation 1:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \qquad (1)$$

In Equation 1, Q, K, and V are the Query, Key, and Value matrices. Here, K is the key vector size; the Softmax function sets a weight for each part of the image and shows how much this weight is compared to the other parts.

Vision Transformers are particularly effective for medical image analysis because they capture both local image features and long-range spatial relationships through self-attention mechanisms. This capability enables accurate identification of subtle morphological changes associated with malaria-infected erythrocytes. The general diagram of the vision transformer model is given in Figure 2.

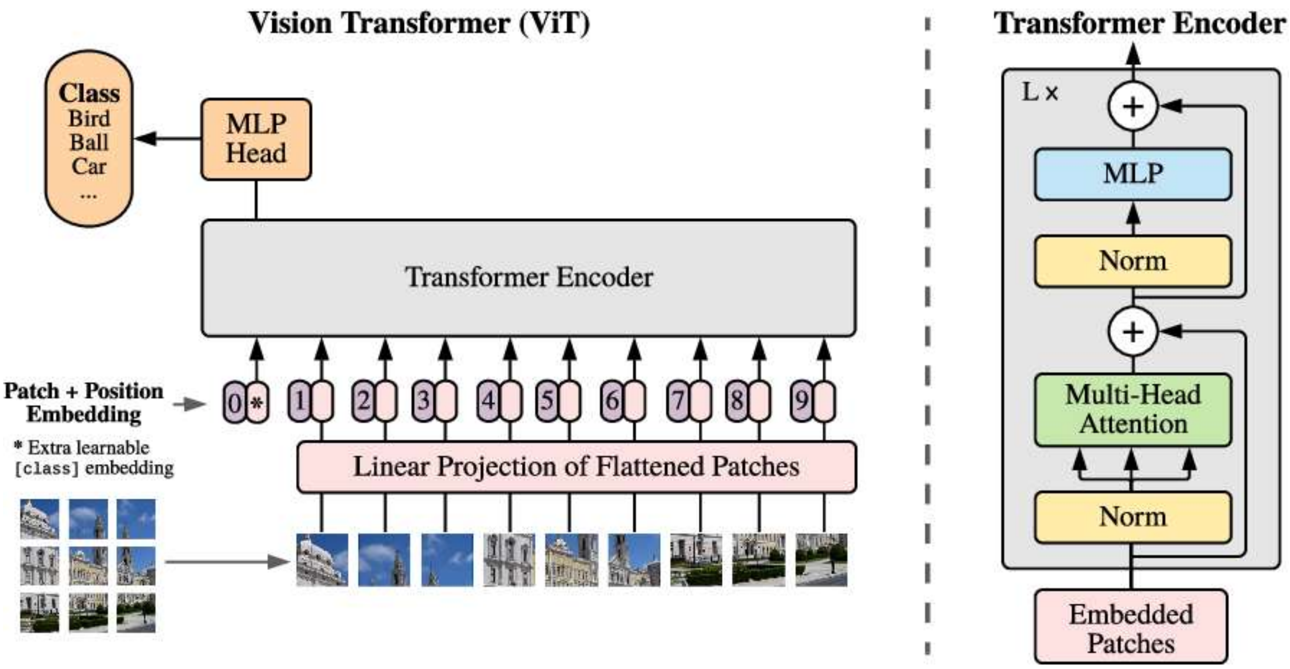

Figure 2. The base vision transformer model [32].

2.5. Base Vision Transformer (BaseViT)

ViT architecture is one of the standard applications for image representations intended to learn images from large-volume datasets. The encoder of the ViT-Base model consists of 12 Transformer layers, each containing 12 self-attention heads, with an embedding dimension of 768. Unlike convolutional neural networks (CNNs), ViT does not rely on convolutional operations for feature extraction; instead, it employs a Transformer-based architecture to capture both local and global dependencies among image regions. The overall structure of the ViT-Base architecture can be described as follows: the input image is first divided into a sequence of fixed-size patches. Each image patch is linearly projected into an embedding vector, and learnable positional embeddings are added to preserve spatial information. These patch embeddings are then fed into the Transformer encoder. Each encoder block consists of a multi-head self-attention (MSA) module and a multilayer perceptron (MLP) block, with layer normalization and residual connections applied to improve training stability and information flow [1], [32]. These components combine to model your image, both local and global connections to learn the presence of malaria, in their definition, infected cells, observing color and texture in terms of subtle differences. The observation made by the BaseViT model can distinguish these cells from healthy cells in a healthy, sensitive way.

2.6. Data-Efficient Image Transformer (DeiT)

The DeiT model, proposed by Touvron and others, is designed for presentation and limited to data vision transformer models taught for an innovative approach offered [33]. The main distinguishing feature of DeiT is the introduction of a distillation token, which is a learnable token designed to facilitate knowledge transfer between a teacher model and a student model. This token is appended to the sequence of patch embeddings along with the class token and is used to transfer information from the teacher network to the student network during training. Unlike conventional Vision Transformers that require large-scale datasets, DeiT employs a knowledge distillation framework that enables the student model to effectively learn from a teacher model, reducing the dependency on extensive training data [34]. Through this approach, DeiT achieves competitive performance while requiring fewer computational resources and smaller datasets. In addition, DeiT incorporates advanced data augmentation strategies during training, which improve model generalization and enhance its ability to extract robust visual representations under different conditions.

2.7. Pyramid Vision Transformer (PVT)

PVT is a hierarchical transformer architecture for multi-dimensional properties to be driven (e.g., object detection and semantic segmentation) predictive tasks. This model has as many stages as the input description size, collecting a variety of rich and high-quality features collector pyramid type hierarchical feature map form. If this structure shows similarities to multi-level Convolutional Neural Networks (CNNs), convolution operations were built without it [35]. Instead, Transformer's focus attention mechanism and various explanations between resolutions place their dependencies on the model for use. The PVT model brings intensive forecasting tasks [36] for both efficient and powerful backbones in image segmentation, such as object detection and tight prediction. The overall diagram of the PVT model is given in Figure 3.

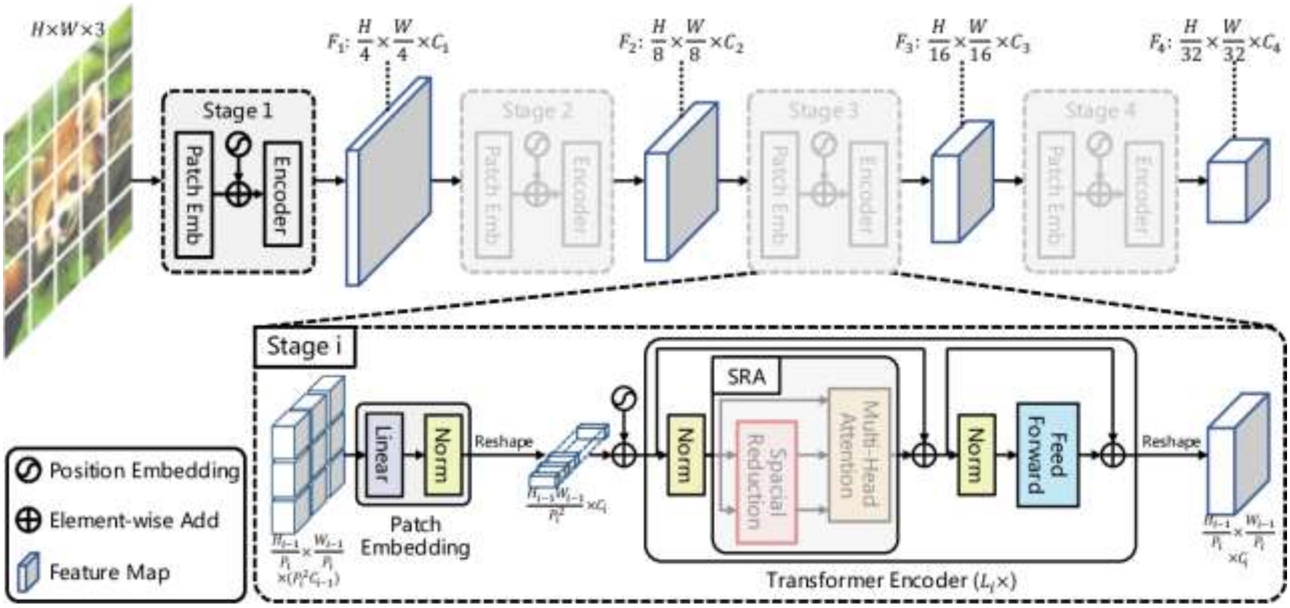

Figure 3. The general diagram of the PVT model [36].

2.8. Swin Transformer

The Swin Transformer model enables a vision transformer architecture to whom it offers, by making sliding windows (shifted windows), balancing efficiency with model complexity, the concept presentation [37]. In this architecture, self-care is not only small and includes other windows located on top of each other, this also reduces significantly, which is calculated to be a computational complexity. The number of neighbouring windows between the exchange of information increases. This mechanism involves neighboring window border markers on windows; they need to be mutually touched in order to create conditions and ultimately provide cross-window communication [38].

2.9. Gradient-Weight Class Activation Mapping (Grad-CAM)

Deep learning models are mainly affected by problems, including a lack of transparency in the process. The Grad-CAM (Gradient-weighted Class Activation Mapping) method is a powerful tool for understanding how to model their decisions. This method emphasizes evolution and layers of attention with related gradients, mapping the model to speech, highlighting the most impact depicting their region. Grad-CAM can also be used for model debugging and reliability evaluation (trust calibration) to be effective in the way of use. This can help researchers and doctors model the predictions behind what is logically better finished. The Grad-CAM heat map is a weighted combination of feature maps followed by a ReLU. Its mathematical formula is given in Equation 2 [39].

$$L_{ij}^{Grad-CAM} = ReLU\left(\sum_k \alpha_k A_{ij}^k\right) \quad (2)$$

Weight Calculation of coefficients (α_k) formula is like:

$$\alpha_k = \frac{1}{Z}\sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k} \quad (3)$$

3. Results and Discussion

3.1. Performance Metrics

Metrics derived from the confusion matrix were used to determine the performance of deep learning models in the detection of malaria-infected cells. These are Accuracy, Sensitivity/Remembering, Specificity, Precise indicator (Precision, and F1 score (F1-Score). These metrics include four possible predicted models to the essentially calculated result: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) [40]. The formulas (balances 4-8) of the metrics derived from the Confusion matrix and their descriptions are given below.

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \quad (4)$$

Here:
TP (True Positive): Correct some positive samples made number (accurate, as open infected cells)
Data TN (True Negative): The number of correctly classified specific negative samples (true), made of clear healthy cells)
DBFP (False Positive): Negative examples with positive error, such as classified cases
DB FN (False Negative): Positive examples with negative error, such as classified cases
Sensitivity (Repeat): Model positive samples (infected cells) to make accurate capability measures precise:

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP+FN}} \quad (5)$$

Specificity: Model negative samples (healthy cells) correctly certain to do ability measuring is an indicator:

$$\text{Specificity} = \frac{\text{TN}}{\text{TN+FP}} \quad (6)$$

Precision: correct predicted positive examples general positive predictions to the number which ratio shows:

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \quad (7)$$

F1 score (F1-Score) Precision and Sensitivity (Recall) harmonic middle like certain is being and this two indicator between balance creates:

$$\text{F1} = 2 \times \frac{\text{Precision}\times\text{Recall}}{\text{Precision+Recall}} \quad (8)$$

Five layers of Cross Verification method was used to perform experimental studies with neural networks. It is a model validation method in which the dataset is divided into five equal parts, one part used for testing and the remaining four for training, and this process is repeated five times to calculate the average performance.

3.2. Experimental Test Results

This study employs eight different deep learning models to detect malaria using peripheral blood images. The performance of four CNN (Convolutional Neural Network) models and four ViT (Vision Transformer) models was evaluated using 5-fold cross-validation. First, the dataset was analyzed using CNN models; the architectures employed included ResNet50, EfficientNet-B0, InceptionResNetV2, and Xception. To achieve optimal results, the training parameters were set as follows: 16 epochs, a mini-batch size of 16, and a learning rate of 1×10^{-4}; these settings were chosen to maximize overall model performance. Experimental results indicate that Xception achieved the best performance among the CNN models, reaching an accuracy of 98.10%. The performance results (confusion matrices) for the CNN models are presented in Table 1 and Figure 4, while the training/validation processes and loss curves are illustrated in Figure 5.

Table 1. Performance results of CNN models.

CNN Models	Accuracy	Sensitivity	Specificity	Precision	F1-Scor
ResNet50	86.19	79.48	95.32	93.18	89.47
EfficientNetb0	94.76	94.79	98.04	94.77	94.48
InceptionResNetV2	96.19	95.22	98.39	97.26	96.07
Xception	98.10	98.02	99.23	98.11	98.01

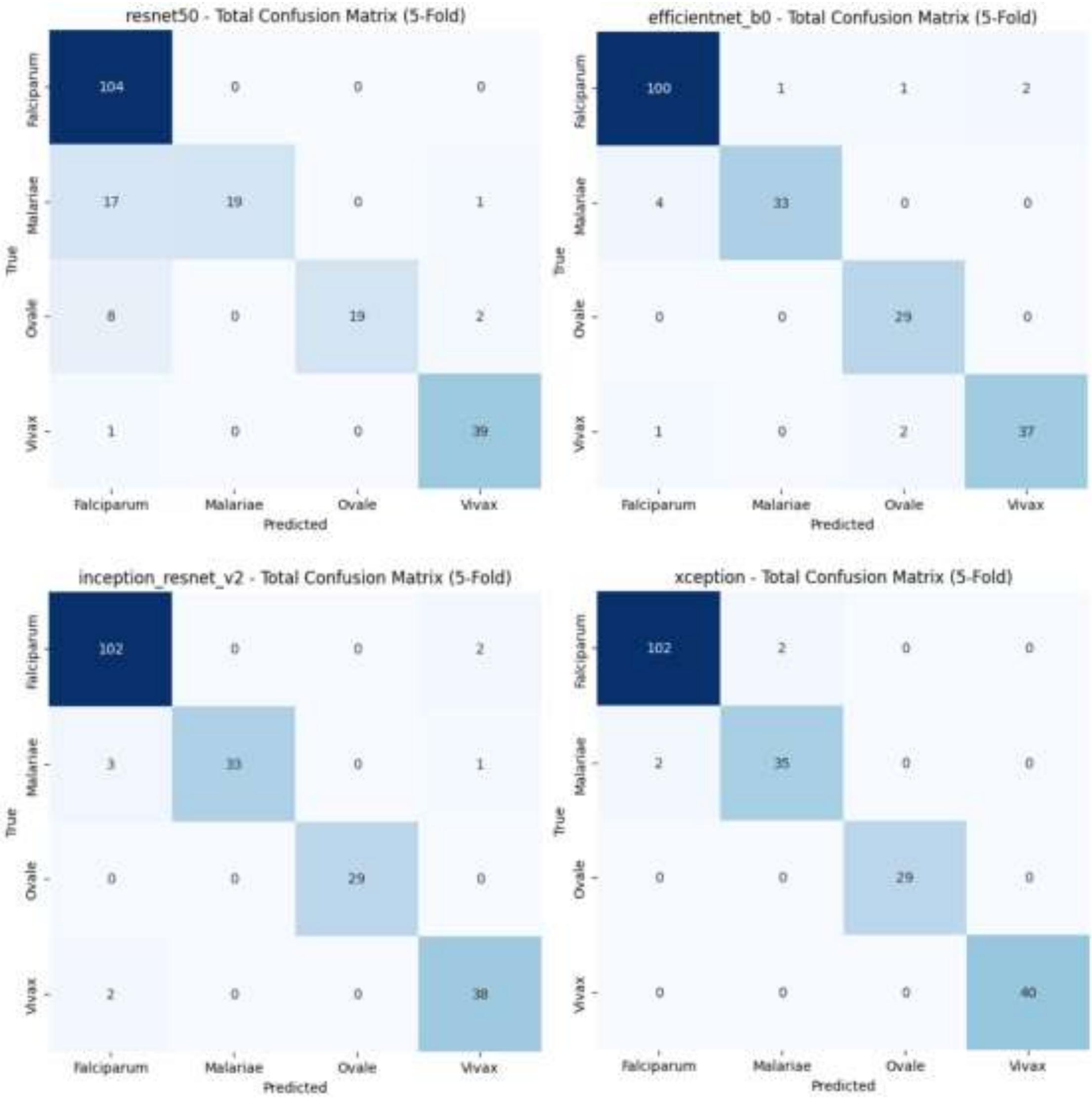

Figure 4. Confusion matrix of CNN algorithms.

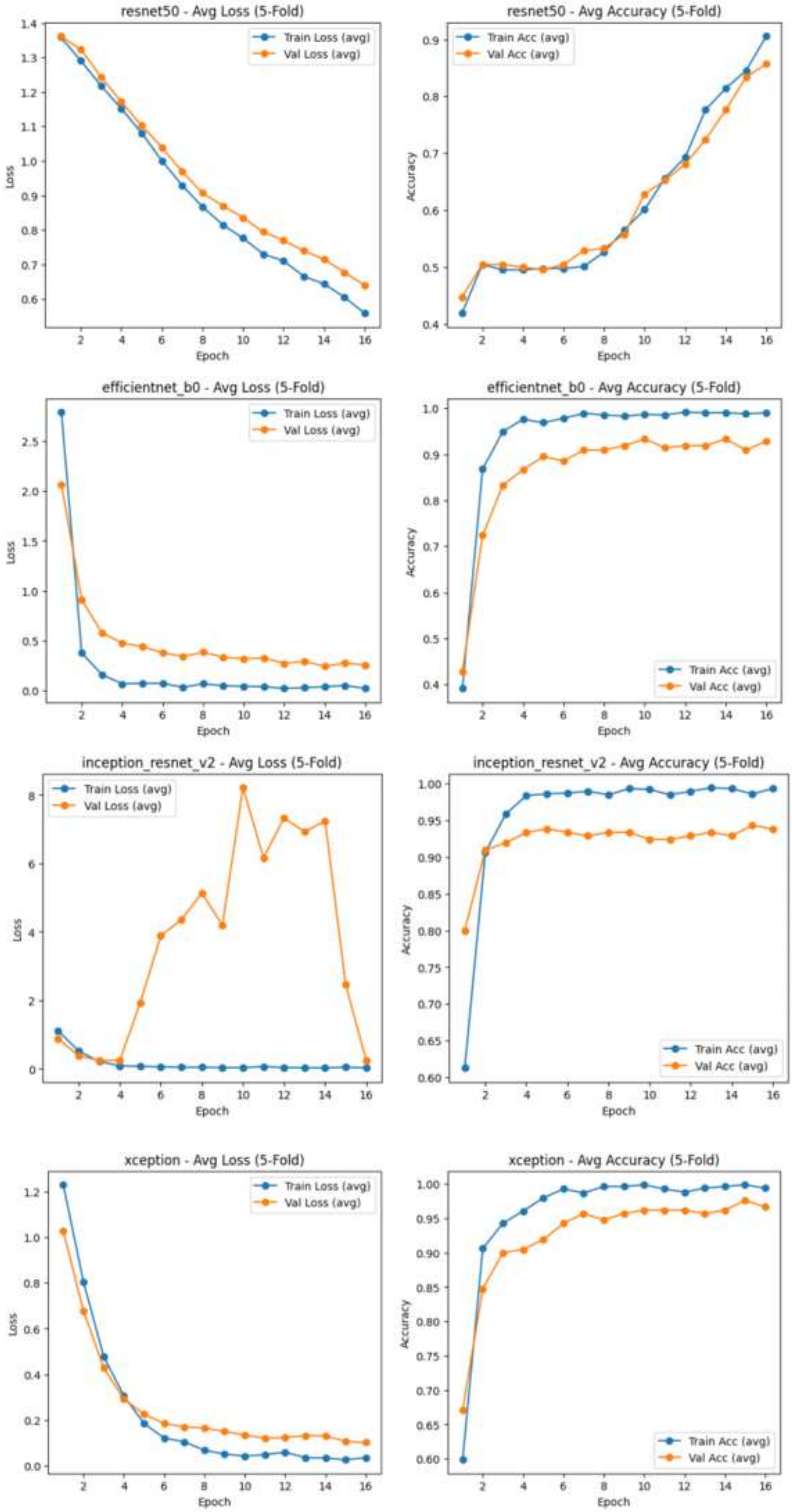

Figure 5. Training and validation accuracy and loss curves.

In the second phase of the experimental work, the dataset was tested using the ViT_base model, Swin Transformer, DeiT, and Pyramid Vision Transformer algorithms. The 5-layer cross-validation method was chosen as the training and test partitioning method for the working dataset. The epoch, mini-batch size, and learning rate for the four ViT models were set to 16, 16, and 1×10^{-4}, respectively, to match those of the CNN models. The optimizer was set to the Adam algorithm, and the loss function was set to cross-entropy loss for hyperparameters. The experimental work resulted in the best performance values with the Swim transformer algorithm and achieved an accuracy rate of 98.10%. Performance results obtained with ViT algorithms are given in Table 2. Confusion matrices are given in Figure 6, and the training of algorithms and the loss graphs are given in Figure 7.

Table 2. Performance results of ViT algorithms.

ViT Models	Accuracy	Sensitivity	Specificity	Precision	F1-Scor
ViT base	96.72	97.31	99.09	97.56	97.42
Swim	98.10	97.73	99.15	98.58	98.13
Deit	96.19	94.89	98.30	97.00	95.86
Pyramid Vit	95.71	93.98	98.18	96.42	94.85

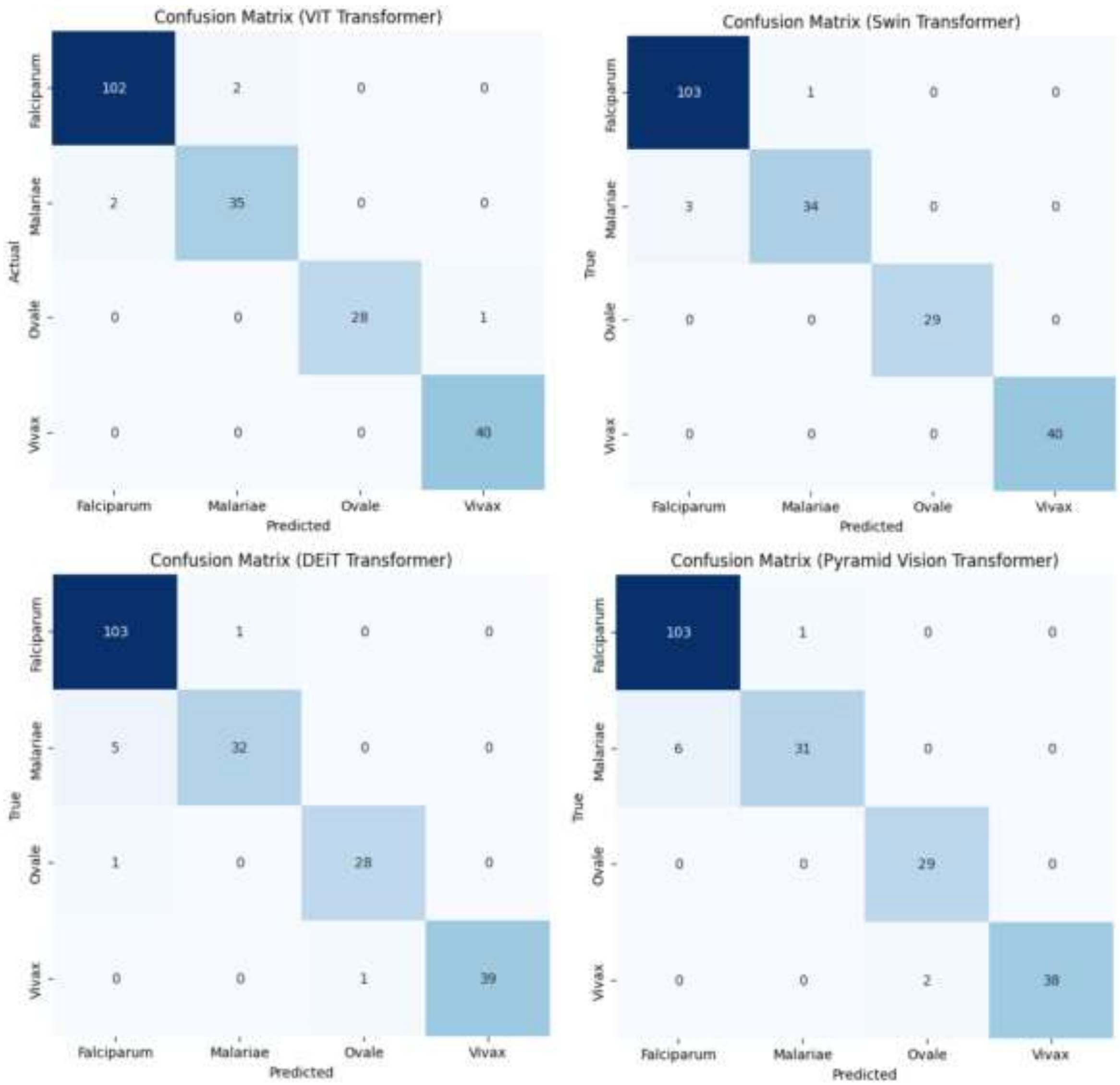

Figure 6. Confusion matrix of ViT algorithms.

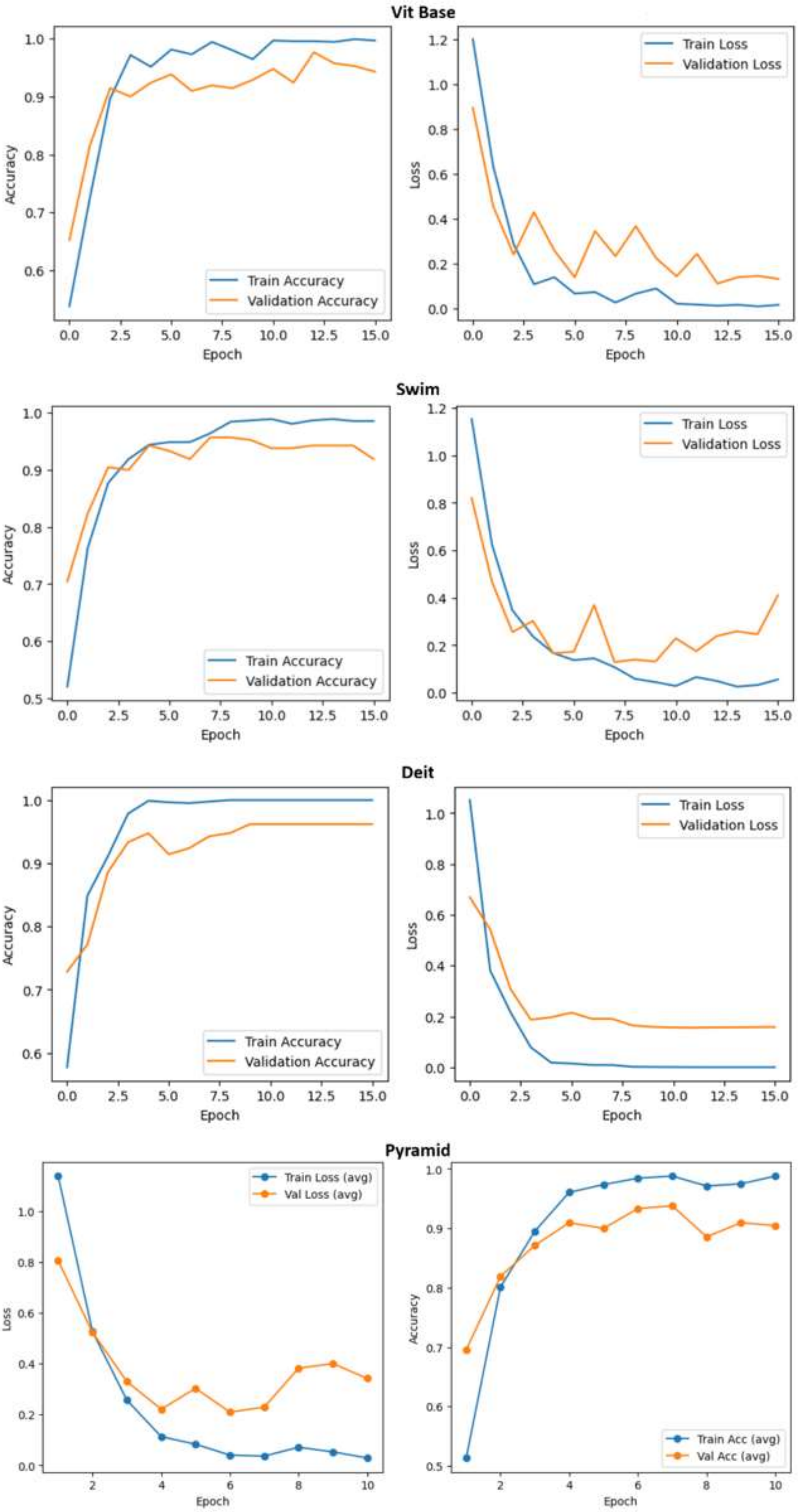

Figure 7. Training and validation loss graph of ViT algorithms.

At the last stage, images were produced using the GRAD-CAM algorithm, an explainable artificial intelligence method that gives the best performance results with the Swin Transformer algorithm. Models can be interpreted with GRAD-CAM images, allowing clinicians to interpret images more easily. Random images generated by the GRAD-CAM algorithm are shown in Figure 8. When examining the images, red lesion areas represent the region that is particularly effective in the decision-making processes of the models. Here, it can be seen that the cells are red in color.

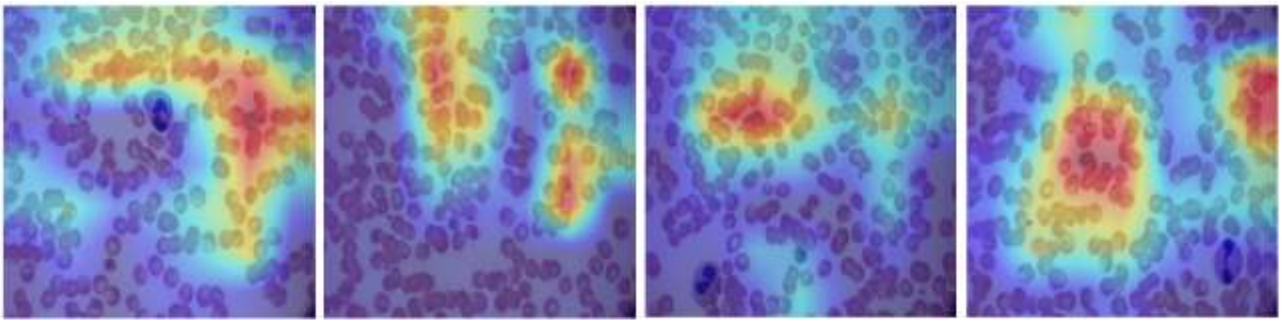

Figure 8. GRAD-CAM images produced with the Swin Transformer.

3.3. Implications for Nanotechnology

Although this study focused on microscopic blood smear image analysis, the proposed explainable AI framework has considerable potential for integration with emerging nanotechnology-enabled diagnostic platforms. Modern malaria diagnostic technologies increasingly employ nanoparticle-based biosensors, microfluidic devices, and portable lab-on-chip systems that generate digital imaging data requiring automated interpretation. Vision Transformer architectures combined with explainable artificial intelligence can provide reliable decision support for these systems by improving classification accuracy while simultaneously increasing clinical interpretability. Therefore, integrating advanced AI algorithms with nanotechnology-based diagnostic platforms represents a promising direction for future malaria diagnostics.

4. Conclusion

Malaria remains a serious health problem, especially for communities living in developing countries, and is one of the most important parasitic infections that threaten public health globally. Diagnostic methods using microscopes, still considered the gold standard nowadays, have certain limitations due to the need for specialist staff, the time required, and the associated labor intensity. This study is aimed at improving the accuracy of malaria detection and automating the process. For this purpose, up-to-date deep learning techniques such as convolutional neural network architectures (ResNet50, EfficientNetB0, InceptionResNetV2, and Xception) and Vision Transformer-based approaches (ViT-Base, Swin Transformer, DeiT, and PVT) were used. This study employed 5-fold cross-validation to train and evaluate models using a dataset of microscopic blood images covering four different types of malaria. Among the convolutional neural network models, Xception achieved the best performance with an accuracy of 98.10%, while the Swin Transformer demonstrated equally excellent performance among the Transformer-based models. Furthermore, the study utilized Gradient-weighted Class Activation Mapping (Grad-CAM) to elucidate the models' decision-making mechanisms, thereby enhancing their reliability for clinical application. This technique enabled researchers to visually demonstrate the specific image regions the models focused on when making diagnostic decisions. The research results reveal that using Transformer-based deep learning architectures in conjunction with explainable artificial intelligence methods significantly improves both classification success and model transparency. These developments enable automated malaria diagnostic systems to be used more reliably in clinical settings and integrated into real-life applications.

At the same time, nanotechnology represents an innovative approach with significant potential for improving the diagnosis and treatment of malaria. Nanomaterial-based delivery systems can enhance the targeted transport of antimalarial drugs to infected cells, improve drug bioavailability, reduce systemic side effects, and increase therapeutic efficacy. In addition, nanotechnology-based platforms contribute to the development of early diagnostic tools and next-generation vaccine strategies. Although many of these technologies remain at the preclinical or clinical research stages, they hold promise for future applications in malaria prevention and treatment. In particular, targeted drug delivery systems based on liposomes, polymeric nanoparticles, and lipid-based nanocarriers enable more efficient delivery of antimalarial agents to infected hepatocytes and erythrocytes. By improving drug accumulation at infection sites while minimizing toxicity to healthy tissues, these nanocarrier systems represent a promising strategy for enhancing malaria therapy.

Future research should investigate the integration of explainable Vision Transformer architectures with nanotechnology-enabled diagnostic systems, including nano-biosensors, plasmonic sensing platforms, microfluidic lab-on-chip devices, and portable digital microscopy. Such multidisciplinary approaches have the potential to improve point-of-care malaria diagnosis by increasing diagnostic accuracy, interpretability, and clinical reliability while facilitating the deployment of intelligent nanodiagnostic technologies in resource-limited settings.

Author Contributions

All authors contributed to the conception and development of the study, participated in the preparation and revision of the manuscript, and approved the final version of the manuscript.

Conflict of Interest

The authors declare that they have no conflicts of interest related to this work.

Funding

This research received no external funding.

Acknowledgment

The authors would like to thank their institutions and colleagues for their support and assistance during the preparation of this study.

Abbreviations

Natural Language Processing (NLP), Gradient-weighted Class Activation Mapping (Grad-CAM), World Health Organization (WHO), Red Blood Cells (RBC), Polymerase Chain Reaction (PCR), Rapid Diagnostic Tests (RDT), Computer Vision (CV), Deep Learning (DL), Convolutional Neural Networks (CNNs), Vision Transformer (ViT), Data Efficient Image Transformer (DeiT), Pyramid Vision Transformer (PVT), Base Vision Transformer (BaseViT), Multi-Head Self-Attention (MSA), Multilayer Perceptron (MLP), True Positive (TP), True Negative (TN), False Positive (FP), False Negative (FN).

References

[1] World Health Organization. World malaria report 2024. Geneva: World Health Organization; 2024. [cited 2025 July 08]. Available from: https://www.who.int/teams/global-malaria-programme/reports/world-malaria-report-2024

[2] Yang F, Poostchi M, Yu H, Zhou Z, Silamut K, Yu J, Maude RJ, Jaeger S, Antani S. Deep learning for smartphone-based malaria parasite detection in thick blood smears. IEEE journal of biomedical and health informatics. 2019 Sep 23;24(5):1427-38. https://doi.org/10.1109/jbhi.2019.2939121

[3] Fong Amaris WM, Martinez C, Cortés-Cortés LJ, Suárez DR. Image features for quality analysis of thick blood smears employed in malaria diagnosis. Malaria journal. 2022 Mar 5;21(1):74. https://doi.org/10.1186/s12936-022-04064-2

[4] Sukumarran D, Hasikin K, Khairuddin AS, Ngui R, Sulaiman WY, Vythilingam I, Divis PC. An optimised YOLOv4 deep learning model for efficient malarial cell detection in thin blood smear images. Parasites & Vectors. 2024 Apr 16;17(1):188. https://doi.org/10.1186/s13071-024-06215-7

[5] Başaran E, Cömert Z, Çelik Y, Velappan S, Toğaçar M. Determination of tympanic membrane region in the middle ear otoscope images with convolutional neural network based YOLO method. Dokuz Eylül Üniversitesi Mühendislik Fakültesi Fen ve Mühendislik Dergisi. 2020 Sep 22;22(66):919-28. https://doi.org/10.21205/deufmd.2020226625

[6] Maturana CR, De Oliveira AD, Nadal S, Bilalli B, Serrat FZ, Soley ME, Igual ES, Bosch M, Lluch AV, Abelló A, López-Codina D. Advances and challenges in automated malaria diagnosis using digital microscopy imaging with artificial intelligence tools: A review. Frontiers in Microbiology. 2022 Nov 15;13:1006659. https://doi.org/10.3389/fmicb.2022.1006659

[7] Goceri E. Deep learning-based classification of facial dermatological disorders. Computers in Biology and Medicine. 2021 Jan 1;128:104118. https://doi.org/10.1016/j.compbiomed.2020.104118

[8] Ahamed MF, Nahiduzzaman M, Mahmud G, Shafi FB, Ayari MA, Khandakar A, Abdullah-Al-Wadud M, Islam SR. Improving Malaria diagnosis through interpretable customized CNNs architectures. Scientific Reports. 2025 Feb 22;15(1):6484. https://doi.org/10.1038/s41598-025-90851-1

[9] Ahishakiye E, Kanobe F, Taremwa D, Nantongo BA, Nkalubo L, Ahimbisibwe S. Enhancing malaria detection and classification using convolutional neural networks-vision transformer architecture. Discover Applied Sciences. 2025 Jun 6;7(6):612. https://doi.org/10.1007/s42452-025-06704-z

[10] Tan D, Liang X. Multiclass malaria parasite recognition based on transformer models and a generative adversarial network. Scientific Reports. 2023 Oct 10;13(1):17136. https://doi.org/10.1038/s41598-023-44297-y

[11] Islam MR, Nahiduzzaman M, Goni MO, Sayeed A, Anower MS, Ahsan M, Haider J. Explainable transformer-based deep learning model for the detection of malaria parasites from blood cell images. Sensors. 2022 Jun 8;22(12):4358. https://doi.org/10.3390/s22124358

[12] Baptista V, Peng WK, Minas G, Veiga MI, Catarino SO. Review of microdevices for hemozoin-based malaria detection. Biosensors. 2022 Feb 11;12(2):110

[13] Rajaraman S, Antani SK, Poostchi M, Silamut K, Hossain MA, Maude RJ, Jaeger S, Thoma GR. Pre-trained convolutional neural networks as feature extractors toward improved malaria parasite detection in thin blood smear images. PeerJ. 2018 Apr 16;6:e4568. https://doi.org/10.7717/peerj.4568

[14] Mujahid M, Rustam F, Shafique R, Montero EC, Alvarado ES, de la Torre Diez I, Ashraf I. Efficient deep learning-based approach for malaria detection using red blood cell smears. Scientific Reports. 2024 Jun 10;14(1):13249. https://doi.org/10.1038/s41598-024-63831-0

[15] Çalışkan A. Diagnosis of malaria disease by integrating chi-square feature selection algorithm with convolutional neural networks and autoencoder network. Transactions of the Institute of Measurement and Control. 2023 Mar;45(5):975-85. https://doi.org/10.1177/01423312221147335

[16] Kaboré KK, Guel D, Somda FH. Advancements in Deep Learning for Malaria Detection: A Comprehensive Overview. International Journal of Advanced Computer Science & Applications. 2025 Jun 1;16(6):57. https://doi.org/10.14569/ijacsa.2025.0160608

[17] Ramos-Briceño DA, Flammia-D'Aleo A, Fernández-López G, Carrión-Nessi FS, Forero-Peña DA. Deep learning-based malaria parasite detection: convolutional neural networks model for accurate species identification of Plasmodium falciparum and Plasmodium vivax. Scientific Reports. 2025 Jan 30;15(1):3746. https://doi.org/10.1038/s41598-025-87979-5

[18] Azad R, Kazerouni A, Heidari M, Aghdam EK, Molaei A, Jia Y, Jose A, Roy R, Merhof D. Advances in medical image analysis with vision transformers: a comprehensive review. Medical image analysis. 2024 Jan 1;91:103000. https://doi.org/10.1016/j.media.2023.103000

[19] Shamshad F, Khan S, Zamir SW, Khan MH, Hayat M, Khan FS, Fu H. Transformers in medical imaging: A survey. Medical Image Analysis. 2023 Aug 1;88:102802. https://doi.org/10.1016/j.media.2023.102802

[20] Halder A, Gharami S, Sadhu P, Singh PK, Woźniak M, Ijaz MF. Implementing vision transformer for classifying 2D biomedical images. Scientific Reports. 2024 May 31;14(1):12567. https://doi.org/10.1038/s41598-024-63094-9

[21] Surendar P, Basha CH, Nabi MS, Kavitha L, Vijayakumar P. Multi-modal deep learning for malaria diagnosis: integrating CNNs and vision transformers for enhanced parasite detection. In2025 Third International Conference on Augmented Intelligence and Sustainable Systems (ICAISS) 2025 May 21 (pp. 570-578). IEEE. https://doi.org/10.1109/icaiss61471.2025.11042087

[22] Awe OO, Mwangi PN, Goudoungou SK, Esho RV, Oyejide OS. Explainable AI for enhanced accuracy in malaria diagnosis using ensemble machine learning models. BMC Medical Informatics and Decision Making. 2025 Apr 11;25(1):162. https://doi.org/10.1186/s12911-025-02874-3
[23] Gao Y, Liu J, Li W, Hou M, Li Y, Zhao H. Augmented grad-cam++: Super-resolution saliency maps for visual interpretation of deep neural network. Electronics. 2023 Nov 30;12(23):4846. https://doi.org/10.3390/electronics12234846
[24] Obeagu EI, Okoroiwu GI, Ubosi NI, Obeagu GU, Onohuean H, Muhammad T, Adias TC. Revolution in malaria detection: unveiling current breakthroughs and tomorrow's possibilities in biomarker innovation. Annals of Medicine and Surgery. 2024 Jul 17;86(10):5859
[25] Loddo A, Di Ruberto C, Kocher M, Prod'Hom G. MP-IDB: the malaria parasite image database for image processing and analysis. InSipaim–Miccai Biomedical Workshop 2018 Sep 20 (pp. 57-65). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-13835-6_7
[26] Başaran E, Cömert Z, Çelik Y. Convolutional neural network approach for automatic tympanic membrane detection and classification. Biomedical Signal Processing and Control. 2020 Feb 1;56:101734. https://doi.org/10.1016/j.bspc.2019.101734
[27] Karadeniz AT, Çelik Y, Başaran E. Classification of walnut varieties obtained from walnut leaf images by the recommended residual block-based CNN model. European Food Research and Technology. 2023 Mar;249(3):727-38. https://doi.org/10.1007/s00217-022-04168-8
[28] Mahesh TR, Gupta M, Anupama TA, Geman O. An XAI-enhanced efficientNetB0 framework for precision brain tumor detection in MRI imaging. Journal of Neuroscience Methods. 2024 Oct 1;410:110227. https://doi.org/10.1016/j.jneumeth.2024.110227
[29] Peng C, Liu Y, Yuan X, Chen Q. Research of image recognition method based on enhanced inception-ResNet-V2. Multimedia Tools and Applications. 2022 Oct;81(24):34345-65. https://doi.org/10.1007/s11042-022-12387-0
[30] Uzun MZ, Başaran E, Çelik Y. Xception Derin Öğrenme Modeli ve Gabor Filtreleri ile ÇDÖÖE-DVM Algoritması Kullanılarak Mikro İfadelerin Tanınması. Journal of the Institute of Science and Technology. 2023 Jan 12;13(4):2339-52. https://doi.org/10.21597/jist.1252556
[31] Liu Z, Mao H, Wu CY, Feichtenhofer C, Darrell T, Xie S. A convnet for the 2020s. In2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR) 2022 Jun 18 (pp. 11966-11976). IEEE. https://doi.org/10.1109/cvpr52688.2022.01167
[32] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929. 2020 Oct 22. https://doi.org/10.48550/arXiv.2010.11929
[33] Touvron H, Cord M, Douze M, Massa F, Sablayrolles A, Jégou H. Training data-efficient image transformers & distillation through attention. In: Meila M, Zhang T, editors. Proceedings of the 38th International Conference on Machine Learning (ICML 2021). Proceedings of Machine Learning Research. Vol. 139. PMLR; 2021. p. 10347-57. [cited 2025 March 07]. Available from: https://proceedings.mlr.press/v139/touvron21a
[34] Toğaçar M. Biotransnet: Detection of plant stress through the conversion of biosensor data into RGB channels and combination with transformer networks. Journal of Crop Health. 2025 Oct;77(5):167. https://doi.org/10.1007/s10343-025-01239-0
[35] Zhou Y, Jiang X, Xu G, Yang X, Liu X, Li Z. PVT-SAR: An arbitrarily oriented SAR ship detector with pyramid vision transformer. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. 2022 Nov 14;16:291-305. https://doi.org/10.1109/jstars.2022.3221784
[36] Wang W, Xie E, Li X, Fan DP, Song K, Liang D, Lu T, Luo P, Shao L. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In2021 IEEE/CVF International Conference on Computer Vision (ICCV) 2021 Oct 10 (pp. 548-558). IEEE. https://doi.org/10.1109/iccv48922.2021.00061
[37] Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, Lin S, Guo B. Swin transformer: Hierarchical vision transformer using shifted windows. In2021 IEEE/CVF International Conference on Computer Vision (ICCV) 2021 Oct 10 (pp. 9992-10002). Ieee. https://doi.org/10.1109/iccv48922.2021.00986
[38] Wei C, Ren S, Guo K, Hu H, Liang J. High-resolution Swin transformer for automatic medical image segmentation. Sensors. 2023 Mar 24;23(7):3420. https://doi.org/10.3390/s23073420

[39] Selvaraju RR, Das A, Vedantam R, Cogswell M, Parikh D, Batra D. Grad-CAM: Why did you say that?. arxiv preprint arxiv:1611.07450. 2016 Nov 22. https://doi.org/10.48550/arXiv.1611.07450
[40] Başaran E, Çelik G, Toğaçar M. Regionally focused neural-coder model designed for the diagnosis of acute lymphoblastic leukemia disease. Measurement. 2025 Dec 1;256:118176. https://doi.org/10.1016/j.measurement.2025.118176

Journal of Nanotechnology and Innovative Materials

Research Article

Effect of Tunnelling on the Thermoelectric Efficiency of Bulk Nanostructured $Bi_{0.5}Sb_{1.5}Te_3$ Materials

Sayyara A. Nabieva[1], Durdana F. Rustamova[2], Ogtay B. Taghiyev[1], Vafa Hajiyeva[1], and Gulnara Huseynova[1]

[1]Institute of Physics, Ministry of Science and Education, 131 H. Javid Avenue, AZ1143 Baku, Azerbaijan
[2]Department of Mechanics and Mathematics, School of Advanced Technologies and Innovation Engineering, Western Caspian University, Ahmad Rajabli 17A, III Parallel, AZ1001 Baku, Azerbaijan

Received: 28.02.2026 Accepted: 20.07.2026 Published: 31.07.2026

https://doi.org/10.54414/QVBK6661

Abstract

This work investigates the influence of electron tunnelling on the kinetic and thermoelectric properties of nanostructured bulk materials based on $Bi_{0.5}Sb_{1.5}Te_3$. The study focuses on systems where electric transport occurs through a network of semiconducting grains separated by narrow dielectric barriers of nanometer-scale thickness. In this regime, the dominant conduction mechanism is tunnelling, and phonon heat transfer across the barriers can be neglected. Numerical evaluation of the electrical conductivity, thermal conductivity, Seebeck coefficient, and the effective thermoelectric figure of merit (ZT) demonstrates that tunnelling can significantly enhance the thermoelectric voltage and lead to $ZT \approx 1\text{–}2$ at room temperature, even though the overall electrical conductivity remains relatively low. These results suggest that properly engineered tunnelling contacts can serve as an efficient mechanism for improving energy conversion in bulk thermoelectric composites.

Keywords: thermoelectric materials, $Bi_{0.5}Sb_{1.5}Te_3$, electron tunnelling, nanostructured composites, Seebeck coefficient, thermal conductivity

1. Introduction

Thermoelectric materials based on Bi–Sb–Te compounds remain among the most efficient systems for room-temperature energy conversion [1], [2], [3]. In recent years, nanostructuring has emerged as a way to enhance thermoelectric performance by suppressing phonon transport, modifying carrier scattering, and altering the electronic density of states [1], [3], [4]. One promising route involves constructing bulk nanocomposites in which semiconductor grains are embedded within a dielectric matrix, forming a network of nanoscale tunnelling junctions [5], [6], [7]. Our system works like this: a network of semiconducting grains, each separated from its neighbours by very narrow dielectric barriers only nanometres thick [8], [9], [10].

The overall transport properties of such a system are governed by both the internal conductivity of the grains and the transmission probability of the tunnelling barriers [5], [6], [11]. For typical Bi–Sb–Te systems, the barrier height is close to 0.8 eV [2], [11].

When the characteristic dimensions of the tunnelling region are shorter than the electron mean free path, carriers pass through the barrier ballistically, eliminating ohmic losses typically associated with scattering [5], [11], [12]. This quantum-mechanical contribution significantly modifies the effective transport coefficients of the composite material and may enhance its energy conversion efficiency.

2. Theoretical Model and Calculation Method

The thermoelectric efficiency of nanostructured $Bi_{0.5}Sb_{1.5}Te_3$ composites was investigated assuming that nanometre-scale dielectric barriers separate semiconductor grains. In such systems, charge transport occurs through tunnelling contacts formed between adjacent grains [5], [6], [11].

The transport mechanism through the tunnelling barrier is assumed to be ballistic, meaning that the barrier thickness is smaller than the electron free path. Under these conditions, electrons pass through the barrier without scattering, and the transmission probability strongly depends on the barrier thickness and height [5], [11].

The equation for the local electrical conductivity (1.1) is expressed as:

$$\sigma_M(x) = \frac{x+\delta}{\frac{x}{\sigma_b(x)} + \frac{\delta}{\sigma} + \frac{(\alpha_b(x)-\alpha)^2}{\delta \cdot \kappa_b(x) + x\kappa}\,\delta x T} \tag{1.1}$$

Here:
σ_M (x) – local electrical conductivity;
x – tunneling barrier thickness;
δ – characteristic grain size;
σ_b (x) – electrical conductivity of the tunneling barrier;
σ – electrical conductivity of the semiconductor grain;
α_b (x) – barrier Seebeck coefficient;
α – grain Seebeck coefficient;
κ_b (x) – barrier thermal conductivity;
κ - grain thermal conductivity.

The equation for local thermal conductivity (1.2) is determined by:

$$\kappa_M(x) = \frac{x+\delta}{\frac{x}{\kappa_b(x)} + \frac{\delta}{\kappa}} \tag{1.2}$$

The equation for the local Seebeck coefficient (1.3) is calculated as:

$$\alpha_M(x) = \frac{\frac{\alpha_b(x)}{\kappa_b(x)}x + \frac{\alpha}{\kappa}\delta}{\frac{x}{\kappa_b(x)} + \frac{\delta}{\kappa}} \tag{1.3}$$

The effective electrical conductivity and thermal conductivity are determined numerically from the integral equations.

The effective electrical conductivity equation (1.4) is expressed as:

$$\int_{x_{min}}^{x_{\max}} \frac{\sigma_e - \sigma_M(x)}{2\sigma_e + \sigma_M(x)} D(x)dx = 0 \tag{1.4}$$

D (x) – distribution function of barrier thicknesses.

The effective thermal conductivity equation (1.5) is determined by:

$$\int_{x_{\min}}^{x_{\max}} \frac{\kappa_e - \kappa_M(x)}{2\kappa_e + \kappa_M(x)} D(x)dx = 0 \quad (1.5)$$

The effective Seebeck coefficient equation (1.6) is calculated as:

$$\alpha_e = \frac{\int_{x_{\min}}^{x_{\max}} \frac{\alpha_M(x)\sigma_M(x)}{\Delta_0(x)} D(x)dx}{\int_{x_{\min}}^{x_{\max}} \frac{\sigma_M(x)}{\Delta_0(x)} D(x)dx} \quad (1.6)$$

After determining the effective electrical conductivity σ_e, thermal conductivity κ_e, and Seebeck coefficient α_e, the thermoelectric figure of merit was calculated using (1.7):

$$ZT = \frac{\sigma\alpha^2 T}{\kappa} \quad (1.7)$$

σ_e – effective electrical conductivity;
α_e – effective Seebeck coefficient;
κ_e – effective thermal conductivity;
T – absolute temperature.

For numerical calculations, the tunnelling barrier height for $Bi_{0.5}Sb_{1.5}Te_3$ was assumed to be 0.8 eV. The calculations were performed at T = 300 K using Mathematica software. The experimentally determined transport parameters for bulk $Bi_{0.5}Sb_{1.5}Te_3$ used in the model were:

$\sigma = 7.0 \times 10^4\ \Omega^{-1}\,m^{-1}$
$k = 1.02$ W/(m·K)
$\alpha = 200$ μV/K

The thermal conductivity of the tunnelling barrier was estimated using the Lorenz/Wiedemann–Franz relation (1.8) [13].

$$\kappa_b = TL_b\sigma_b \quad (1.8)$$

L – Lorenz number;
σ_b – barrier conductivity;
T – temperature.

The local electrical conductivity $\sigma_a(x)$ (1.9) is given by:

$$\sigma_a(x) = (x + \delta) / [x/\sigma\_b(x) + \delta/\sigma + ((\alpha\_b(x) - \alpha)^2/ (\delta\cdot\kappa\_b(x) + x\cdot\kappa)) \cdot \delta\cdot x\cdot T] \quad (1.9)$$

x – tunnelling barrier thickness;
δ – characteristic grain size;
σ_b(x) – conductivity of the barrier;
σ – conductivity of the grain;
α_b(x) – Seebeck coefficient of the barrier;

α – Seebeck coefficient of the grain;
κ_b(x) – thermal conductivity of the barrier;
κ – thermal conductivity of the grain.

The local thermal conductivity $\kappa_a(x)$ (1.10) is:

$$\kappa_a(x) = (x + \delta) / (x/\kappa\_b(x) + \delta/\kappa) \qquad (1.10)$$

The local Seebeck coefficient $\alpha_a(x)$ (1.11) is:

$$\alpha_a(x) = [(\alpha\_b(x)/\kappa\_b(x))\, x + (\alpha/\kappa)\cdot\delta] / [x/\kappa\_b(x) + \delta/\kappa] \qquad (1.11)$$

To get the effective electrical and thermal conductivities of the whole composite, we solved these integral equations (1.12), (1.13):

$$\int [(\sigma_e - \sigma_a(x)) / (2\sigma_e + \sigma_a(x))] \cdot D(x)\, dx = 0 \qquad (1.12)$$

$$\int [(\kappa_e - \kappa_a(x)) / (2\kappa_e + \kappa_a(x))] \cdot D(x)\, dx = 0 \qquad (1.13)$$

D(x) – distribution of barrier thicknesses.

The effective Seebeck coefficient (1.14), (1.15) came from:

$$\alpha_e = \int [(\alpha_a(x)\cdot\sigma_a(x)) / \Delta_0(x)] \cdot D(x)\, dx \, / \int [\sigma_a(x) / \Delta_0(x)] \cdot D(x)\, dx \qquad (1.14)$$

$$\text{with} \quad \Delta_0(x) = (2\sigma_e + \sigma_a(x)) \cdot (2\kappa_e + \kappa_a(x)) \qquad (1.15)$$

Once we had σ_e, κ_e, and α_e, we calculated the figure of merit (1.16):

$$ZT = (\sigma_e \cdot \alpha_e^2 \cdot T) / \kappa_e \qquad (1.16)$$

3. Results and Discussion

Figure 1 shows the dependence of the effective electrical conductivity on the minimum thickness of the tunnelling barrier.

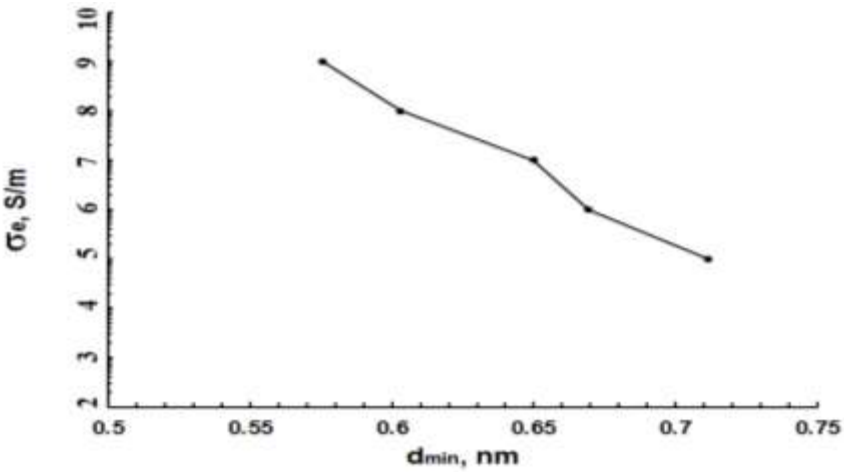

Figure 1. Dependence of the effective electrical conductivity on the minimum tunnelling barrier thickness [6].

The conductivity decreases exponentially as the barrier thickness increases. This behaviour is characteristic of quantum tunnelling transport because the tunnelling probability rapidly decreases with increasing barrier width [5], [6], [11]. For ultrathin barriers, electrons can pass through the contact with relatively low resistance, whereas thicker barriers strongly suppress carrier transport.

Figure 2 presents the calculated dependence of thermal conductivity on the tunnelling barrier thickness. Similar to electrical conductivity, the thermal conductivity also decreases with increasing barrier thickness. However, the reduction in thermal conductivity is especially important for improving thermoelectric efficiency because phonon transport is suppressed more strongly than electronic transport in the tunnelling region [1], [3], [7].

The comparatively low thermal conductivity of the tunnelling barrier contributes to the enhancement of thermoelectric performance by reducing heat leakage through the composite structure.

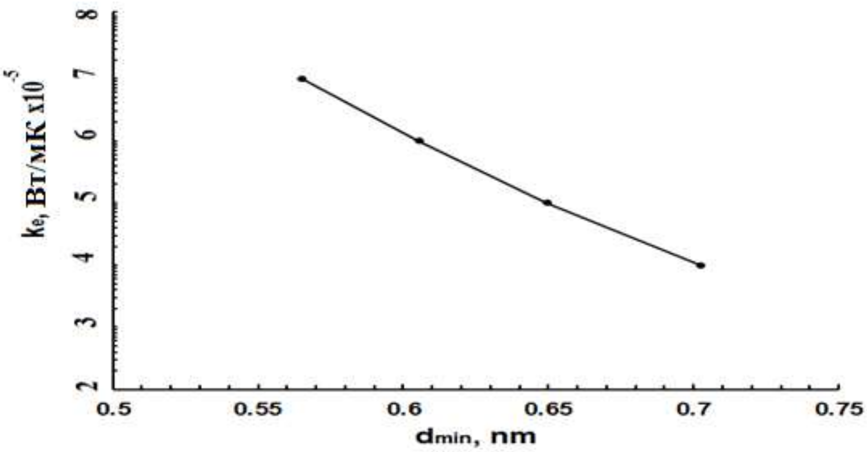

Figure 2. Calculated dependence of thermal conductivity on the minimum tunnelling barrier thickness [6].

Figure 3 illustrates the dependence of the effective Seebeck coefficient on the minimum tunnelling barrier thickness. The Seebeck coefficient increases as the barrier thickness decreases. This effect can be explained by the energy filtering of charge carriers inside the tunnelling barrier [3], [4]. Electrons with higher energy have a greater probability of transmission through the barrier, which increases the average carrier energy and therefore enhances the thermoelectric voltage.

The energy filtering effect observed in the model is consistent with previously reported tunnelling and low-dimensional thermoelectric systems [3], [4], [12], [14], [15].

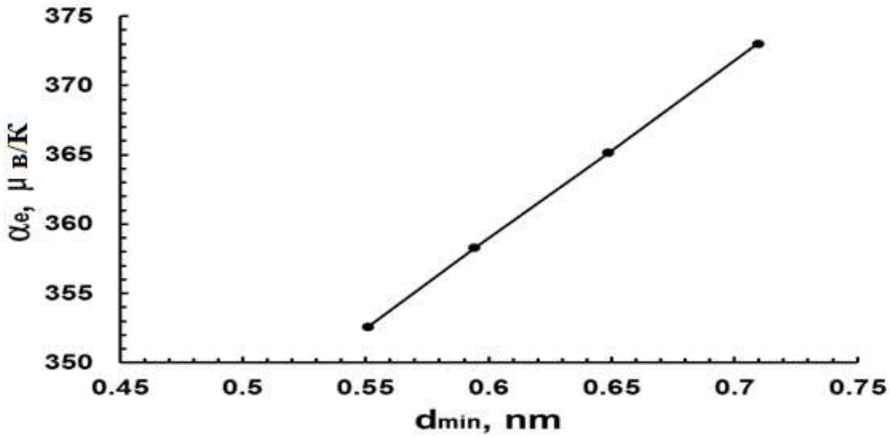

Figure 3. Dependence of the effective Seebeck coefficient on the minimum tunnelling barrier thickness [6].

Figure 4 shows the calculated dependence of the thermoelectric figure of merit *ZT* on the tunnelling barrier thickness. The results indicate that *ZT* increases significantly for ultrathin tunnelling barriers. According to the theoretical model, the effective thermoelectric efficiency may approach values close to $ZT \approx 1\text{–}2$ under idealized tunnelling conditions at room temperature.

The enhancement of thermoelectric efficiency is primarily associated with the strong suppression of thermal conductivity, together with the simultaneous increase in the Seebeck coefficient. Although the electrical conductivity decreases with increasing barrier thickness, the reduction in heat transport dominates, leading to an overall increase in *ZT*.

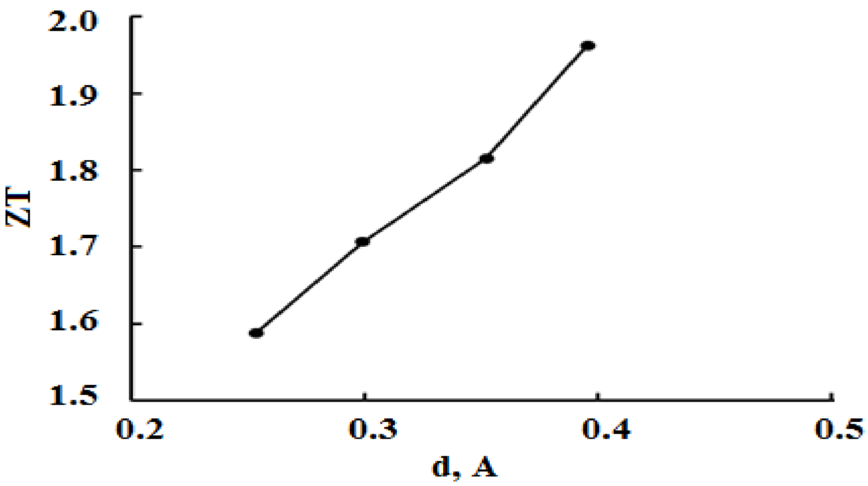

Figure 4. Calculated dependence of the thermoelectric figure of merit *ZT* on the minimum tunnelling barrier thickness [6].

The obtained results demonstrate that the thermoelectric properties of nanostructured $Bi_{0.5}Sb_{1.5}Te_3$ composites are highly sensitive to the geometric parameters of the tunnelling contacts, particularly the barrier thickness. Therefore, precise control of nanoscale barrier formation may provide an effective route for improving thermoelectric energy conversion efficiency.

4. Conclusion

This work demonstrates that tunnelling transport in nanostructured $Bi_{0.5}Sb_{1.5}Te_3$ composites can significantly modify their thermoelectric properties. The presence of nanometre-scale barriers leads to a substantial reduction of thermal conductivity while preserving or enhancing the Seebeck coefficient. As a result, the effective thermoelectric efficiency can reach values of $ZT \approx 1\text{–}2$ at room temperature. The influence of barrier geometry, grain shape, and material distribution is secondary compared to the dominant role of barrier thickness and height. These results confirm that engineering tunnelling junctions is a promising approach for designing next-generation high-efficiency thermoelectric materials. The behaviour predicted by the model is qualitatively consistent with experimental reports on nanostructured Bi_2Te_3-based materials, in which enhancements in ZT have been observed due to grain-boundary scattering and energy filtering [1], [3], [4]. However, the current model uniquely demonstrates that quantum tunnelling alone, without the involvement of phonon engineering techniques, is sufficient to produce $ZT \approx 1\text{–}2$, a significant theoretical insight. Furthermore, the results agree with the predictions of Dresselhaus (2007) regarding the benefits of nanoscale confinement but provide a different physical mechanism applicable to bulk composites [3].

Author Contributions

All authors reviewed, edited, and approved the manuscript.

Conflict of Interest

The authors declare no conflicts of interest.

Funding

This research received no external funding.

Acknowledgment

The authors declare no acknowledgments.

Abbreviations

Thermoelectric Figure of Merit (ZT), Bismuth–Antimony Telluride ($Bi_{0.5}Sb_{1.5}Te_3$).

References

[1] Bulat LP, Osvenskii VB, Parkhomenko YN, Pshenai-Severin DA. Investigation of the possibilities for increasing the thermoelectric figure of merit of nanostructured materials based on Bi2Te3-Sb2Te3 solid solutions. Physics of the Solid State. 2012 Nov;54(11):2165-72. https://doi.org/10.1134/s1063783412110054
[2] Goltsman BM, Kudinov VA, Smirnov IA. Semiconductor thermoelectric materials based on Bi2Te3.
[3] Dresselhaus MS, Chen G, Tang MY, Yang RG, Lee H, Wang DZ, Ren ZF, Fleurial JP, Gogna P. New directions for low-dimensional thermoelectric materials. Advanced Materials. 2007 Apr 20;19(8):1043-53. https://doi.org/10.1002/adma.200600527
[4] Harman TC, Taylor PJ, Walsh MP, LaForge BE. Quantum dot superlattice thermoelectric materials and devices. science. 2002 Sep 27;297(5590):2229-32. https://doi.org/10.1126/science.1072886
[5] Burstein E, Lundqvist S, editors. Tunneling phenomena in solids. New York: Plenum Press; 1969. https://doi.org/10.1007/978-1-4684-1752-4
[6] Bartkowiak M, Mahan GD. Boundary effects in thin-film thermoelectrics. MRS Online Proceedings Library (OPL). 1998 Jan;545:265. https://doi.org/10.1557/proc-545-265
[7] Da Silva LW, Kaviany M. Micro-thermoelectric cooler: interfacial effects on thermal and electrical transport. International Journal of Heat and Mass Transfer. 2004 May 1;47(10-11):2417-35. https://doi.org/10.1016/j.ijheatmasstransfer.2003.11.024
[8] Anatychuk LI, Bulat LP. Semiconductors under extreme temperature conditions. Saint-Petersburg: Nauka Publ.–224 p. 2001.
[9] Babin VP, Gudkin TS, Dashevskii ZM, Dashevskii LD, Dudkin EK. Lordanishvili, EK; Kaidanov, VI; Kolomoets, NV; Narva, OM; Stil'bans, LS. Anisotropic synthetic thermoelements and their maximum capabilities. Sov. Phys. Semicond. 1974;8:478-81.
[10] Snarskii AO. Effective properties of macroscopically inhomogeneous media with restructuring. Ukrainian journal of physics. 2025(70,№ 10):697-703. https://doi.org/10.15407/ujpe70.10.697
[11] Stratton R. Theory of field emission from semiconductors. Physical Review. 1962 Jan 1;125(1):67. https://doi.org/10.1103/physrev.125.67
[12] Hishinuma Y, Geballe TH, Moyzhes BY, Kenny TW. Refrigeration by combined tunneling and thermionic emission in vacuum: Use of nanometer scale design. Applied Physics Letters. 2001 Apr 23;78(17):2572-4. https://doi.org/10.1063/1.1365944
[13] Landau LD, Bell JS, Kearsley MJ, Pitaevskii LP, Lifshitz EM, Sykes JB. Electrodynamics of continuous media. Elsevier; 2013 Oct 22.
[14] Mahan GD. Thermionic refrigeration. Journal of Applied Physics. 1994 Oct;76(7):4362-6. https://doi.org/10.1063/1.357324
[15] Mahan GD. Multilayer thermionic refrigeration. In Eighteenth International Conference on Thermoelectrics. Proceedings, ICT'99 (Cat. No. 99TH8407) 1999 Aug 29 (pp. 584-589). IEEE. https://doi.org/10.1109/ict.1999.843457

Journal of Nanotechnology and Innovative Materials

Review Article

Mechanical Behaviour and Structural Dynamics of Nanomaterials

Farajova S. Sona

Department of Mechanics, Baku State University, Z. Khalilov Street 23, AZ1148 Baku, Azerbaijan

Received: 28.06.2026 Accepted: 18.07.2026 Published: 31.07.2026

https://doi.org/10.54414/MVVP6826

Abstract

Nanomaterials exhibit distinctive mechanical and structural properties that arise from their nanoscale dimensions, high surface-to-volume ratios, and size-dependent interactions. These characteristics influence their elastic modulus, hardness, strength, plasticity, deformation behavior, and dynamic response, distinguishing nanomaterials from their bulk counterparts. This review summarizes the principal classes of nanomaterials and examines the experimental techniques commonly used to characterize their structural and mechanical properties, including atomic force microscopy, scanning and transmission electron microscopy, X-ray diffraction, and Fourier-transform infrared spectroscopy. Particular attention is given to size and shape effects, grain-boundary contributions, surface effects, and the mechanical behavior of nanowires, nanoplates, polymer nanocomposites, carbon-based nanomaterials, and metal nanoparticles. The review also discusses theoretical approaches used to describe nanoscale mechanical behavior, with emphasis on nonlocal elasticity, strain-gradient theories, couple-stress models, and stress-driven integral formulations. Wave propagation and vibration in nanobeams, nanotubes, and multilayer structures are considered in relation to their dynamic behavior and potential applications. Finally, current challenges and future research directions are outlined, emphasizing the development of reliable theoretical and computational frameworks for the design of advanced nano-enabled materials and devices.

Keywords: nanomaterials, mechanical properties, structural dynamics, size effects, nonlocal elasticity, strain-gradient theory, wave propagation, nanostructures, nanomechanics, mechanical characterization

1. Introduction

Over the past two decades, nanomaterials have become one of the most rapidly developing fields of materials science. Materials with structural features on the nanometer scale (approximately 1–100 nm) exhibit physical, chemical, and mechanical properties that differ markedly from their bulk counterparts. This divergence arises primarily from quantum size effects, a sharply increased surface-to-volume ratio, and altered interatomic interactions. As a result, nanomaterials can exhibit distinctive mechanical, electrical, optical, thermal, and catalytic properties compared with their bulk counterparts, making them attractive for applications ranging from medicine and electronics to aerospace engineering and energy storage [1], [2].

The rapid advancement of nanotechnology is driven not only by the discovery of new materials but also by the atomically controlled structuring of existing ones. The resulting nanomaterials exhibit mechanical strength, flexibility, and dynamic response characteristics that cannot be achieved under ordinary bulk conditions, which makes them appealing for both fundamental research and industrial application [2], [3].

The field of nanomaterials has also created an important research domain for applied mathematics and structural mechanics. Classical continuum mechanics may not adequately capture certain size-dependent phenomena observed in nanostructures, particularly when surface, nonlocal, and microstructural effects become significant. This review systematically examines the principal classes of nanomaterials, their mechanical properties, the structural-mechanical approaches used to describe these properties, and their practical fields of application.

2. Classification of Nanomaterials

Nanomaterials are commonly classified according to their chemical composition and morphology. Organic nanomaterials include micelles, dendrimers, polymersomes, and hydrogels, while inorganic nanomaterials consist of metal- and metal-oxide-based particles [1], [4]. Metal nanoparticles, for example, those based on gold, silver, copper, and zinc, display unusual optical, electrical, and catalytic properties owing to their quantum effects and high surface-to-volume ratio.

Carbon-based nanomaterials form a distinct class encompassing fullerenes, carbon nanotubes, graphene, and nanodiamonds. Owing to their sp^2-hybridized carbon structures, these materials exhibit high mechanical strength, electrical conductivity, and thermal conductivity [5]. Composite nanostructures arise from the combination of different classes of nanomaterials and frequently produce synergistic effects that exceed the properties of the individual components. Synthesis methods also significantly affect the ultimate properties of nanomaterials. Laser ablation, sputtering, chemical vapor deposition, the sol-gel method, and green synthesis based on plant extracts are among the most widely used approaches [1], [4]. Careful control of synthesis conditions, including temperature, reaction rate, and precursor concentration, determines particle size, shape, size distribution, and grain-boundary structure, which in turn influence mechanical behavior.

3. Characterization Methods

Accurate assessment of the structural and mechanical properties of nanomaterials relies on a range of modern characterization techniques. Atomic force microscopy (AFM) enables nanometer-precision mapping of the surface topography of nanostructures and is also used in resonant-contact mode to directly measure the elastic modulus. Scanning electron microscopy (SEM) and transmission electron microscopy (TEM) serve to visualize particle morphology, size distribution, and internal crystalline structure, while in-situ mechanical testing modes allow simultaneous recording of load-deformation curves [2], [3].

Ultraviolet-visible (UV–Vis) spectroscopy, Fourier-transform infrared (FTIR) spectroscopy, X-ray diffraction (XRD), and vibrating-sample magnetometry (VSM) are used to characterize the optical properties, functional groups, crystal structure, and magnetic properties of nanomaterials, respectively [1]. The combination of these methods allows researchers to characterize both the static structure and the dynamic mechanical response of a nanomaterial under external loading in a unified way, providing the experimental database necessary for validating theoretical models.

4. Mechanical Properties: General Trends

The mechanical properties of nanomaterials elastic modulus, hardness, strength, and plasticity undergo systematic changes as size decreases. The widely accepted 'smaller is stronger' trend indicates that strength increases as characteristic size decreases, whereas the size-dependence of elastic modulus is more complex and does not follow a single, unambiguous pattern [2], [6]. Some studies report an increase in elastic modulus with decreasing size, while others report a decrease, a divergence stemming from the interplay of surface tension, grain-boundary structure, and the geometric shape of the nanostructure.

In nanocrystalline materials, a decrease in grain size typically leaves the elastic modulus close to the bulk value down to a grain size of roughly 5 nanometers; below this threshold, however, the modulus changes sharply because the volume fraction of grain boundaries begins to dominate the overall structure [2], [3]. This is among the key empirical findings demonstrating the limitations of classical continuum approaches in describing the mechanical behavior of nanomaterials.

Experimental studies on nanowires and nanoplates further show that not only size but also the shape of a nanostructure significantly affects its elastic properties, a factor long neglected in the literature. In-situ electron microscopy tests on gold nanowires have confirmed that the combined effect of shape and size substantially alters the value of the Young's modulus. Similarly, in copper nanowires, an increased surface-to-volume ratio has been shown to link elastic properties directly to the proportion of surface atoms [6].

In polymer nanocomposites, elastic modulus and yield stress are modeled as functions of nanofiller content, temperature, loading rate, and porosity [7]. Such models typically relate porosity to elastic modulus through power-law approaches and allow the reliability of polymer-nanocomposite membranes to be assessed for applications such as water purification and bone tissue engineering.

In carbon-based nanomaterials, particularly multiwalled carbon nanotubes and graphene flakes, mechanical properties also play an important role in interactions with biological media. Atomic force microscopy measurements show that cells exposed to these materials undergo changes in actin cytoskeleton structure, which in turn is reflected in a decrease in the measured Young's modulus of the cell. These findings indicate that the mechanical properties of nanomaterials are relevant not only to structural applications but also to biosafety assessment, since changes in cell mechanics can serve as an indicator of potential cytotoxicity.

In metal nanoparticles, contact mechanics deviates from the classical Hertzian model, since atomic-scale surface steps and structural asymmetry during compression violate the assumptions of continuum models. Molecular dynamics simulations serve as a complementary tool in this area: studies of copper nanowires under tensile load, for example, have shown that elastic properties are directly related to the proportion of surface atoms among the total atoms. Similarly, orbital-free calculations based on density functional theory have shown that in large nanosystems, cohesive energy and bulk modulus gradually approach the known values for bulk material.

5. Structural Mechanics and Wave Propagation Aspects

The inability of classical elasticity theory to fully describe mechanical behavior in nanostructures has directed researchers toward nonlocal continuum models. Eringen's nonlocal elasticity theory is among the most widely used frameworks in this area; according to this theory, the stress at a given point depends not only on the strain at that point but on the strain field of the entire body [8], [9]. This approach has found broad application in modeling the bending, compression, vibration, and wave propagation behavior of nanoplates, nanotubes, and nanolayers.

Alongside nonlocal elasticity theory, strain-gradient models have also been developed, incorporating higher-order gradient terms in addition to the classical strain field [8], [9]. Lim and colleagues combined Eringen's nonlocal law with strain-gradient elasticity to study wave propagation problems in unbounded media, obtaining higher-order differential constitutive equations. Such combined models make it possible to describe both softening and hardening effects simultaneously, a result that purely nonlocal or purely gradient models cannot achieve on their own.

Wave propagation analysis is of particular importance in studying the mechanical properties of nanostructures, since it is computationally efficient and does not require large volumes of experimental data. In Timoshenko- and Rayleigh-type nanobeams, dispersion characteristics and the dependence of phase and group velocity on wavenumber have been studied while accounting for the nonlocal parameter, temperature effects, magnetic fields, and elastic foundations [10]. Stress-driven integral models have eliminated the ill-posed boundary-value problems arising from the Eringen-Wieghardt approach, offering more consistent solution strategies.

Nanotubes and multilayer nanostructures with eccentric (off-center) geometry are of particular importance for studying how initial stresses, geometric asymmetry, and material heterogeneity affect wave dispersion characteristics. Unlike classical concentric cylinder models, this class of problems requires additional mathematical complexity, asymmetric boundary conditions, and additional geometric parameters, making it an ongoing and active research direction in structural mechanics and applied mathematics.

Wave propagation in magneto-electro-elastic (MEE) nanotubes and nanobeams has also been the subject of intensive study in recent years [11]. Because these materials exhibit coupled mechanical, electrical, and magnetic field effects, their dynamic models require combining nonlocal stress theory with modified couple-stress theory. Such coupled models play a central role in the design of smart nano-devices, sensors, and actuators.

Because the analytical solution of wave propagation problems becomes difficult under complex boundary conditions, researchers also turn to semi-analytical and numerical methods. The wave-based method is used to determine the wave characteristics of nanobeams with arbitrary boundary conditions: boundary conditions are incorporated into the construction of the system's total matrix, and the points where the determinant of this matrix vanishes are sought to determine natural frequencies. Similarly, numerical approaches such as the generalized differential quadrature (GDQ) method and the nonlocal operator method offer cost-effective alternatives for complex-geometry and heterogeneous nanostructures where analytical solutions are not feasible, eliminating the need to compute shape functions and their derivatives and thereby improving both accuracy and computational speed.

6. Applications

The practical outcomes of research into the mechanical properties of nanomaterials are evident across a wide range of industrial and medical fields. In medicine, nanoparticles are used for drug delivery, tissue engineering, bioimaging, and biosensing; in these applications, the mechanical stiffness of particles directly determines their interaction with cell membranes and their rate of biodegradation.

In structural engineering, polymer composites reinforced with graphene oxide and carbon nanotubes substantially increase elastic modulus and tensile strength; some studies have reported an increase in elastic modulus of up to 150% with the addition of a small amount of graphene oxide. This property is widely used in the development of lightweight yet high-strength construction materials.

In electronics, nanowires and nanoplates have a broad application spectrum, ranging from flexible displays to high-frequency nano-resonators. Reliable operation of these devices requires precise mathematical modeling of their dynamic behavior, particularly wave dispersion and vibration characteristics, since classical engineering approaches do not provide sufficient accuracy at the nanoscale.

In environmental applications, nanomaterials are used in processes such as pollutant adsorption from water, air purification, and soil remediation, where high surface area enhances adsorption efficiency. In the energy sector, nanostructured electrode materials are being investigated to increase the capacity of batteries and supercapacitors, making the mechanical stability of nanomaterials a critical factor for long-term operation.

In agriculture and the food industry, nanomaterials are used to improve the barrier properties of packaging materials, create antimicrobial coatings, and extend the shelf life of food products. In these applications, the mechanical elasticity and crack resistance of nanolayers are key parameters ensuring product resistance to damage during transport. In catalytic processes, the high surface energy of metal and metal-oxide nanoparticles significantly increases reaction rates, making them valuable across fields ranging from the petrochemical industry to environmental remediation.

In sensor technologies, the combination of mechanical flexibility and electrical conductivity in nanomaterials has enabled the development of wearable devices, biosensor chips, and structural health monitoring systems. In such sensors, the sensitivity of the nanostructure to small deformations, which depends directly on its elastic modulus and wave propagation characteristics, is a primary factor determining measurement accuracy, once again highlighting the connection between theoretical mechanics and applied engineering.

7. Conclusion and Future Directions

Owing to their unique size-dependent mechanical, electrical, and optical properties, nanomaterials have become one of the central research subjects of modern materials science. As shown in this review, the classification of nanomaterials, their synthesis methods, the size-dependence of their mechanical properties, and the mathematical modeling of these properties within the frameworks of nonlocal elasticity and strain-gradient theories constitute closely interrelated research directions.

One of the principal challenges for future research is unifying the various theoretical models, nonlocal, strain-gradient, and couple-stress, within a single, computationally efficient framework. Furthermore, the accurate description of wave dispersion and dynamic behavior in nanostructures with complex geometries, including eccentric and multilayer cylindrical structures, remains an ongoing and active research direction, both theoretically and experimentally. Progress in this area will lay the groundwork for more reliable and effective application of nanotechnologies in medicine, energy, electronics, and structural engineering.

Author Contributions

The author conceived, prepared, reviewed, and approved the final version of the manuscript.

Conflict of Interest

The author declares no conflict of interest.

Funding

This research received no external funding.

Acknowledgment

Not applicable.

Abbreviations

Atomic Force Microscopy (AFM), Scanning Electron Microscopy (SEM), Transmission Electron Microscopy (TEM), Fourier-Transform Infrared Spectroscopy (FTIR), X-Ray Diffraction (XRD), Magneto-Electro-Elastic (MEE), Generalized Differential Quadrature (GDQ), Ultraviolet-Visible spectroscopy (UV–Vis), Vibrating-Sample Magnetometry (VSM).

References

[1] Mekuye B, Abera B. Nanomaterials: An overview of synthesis, classification, characterization, and applications. Nano Select. 2023 Aug;4(8):486-501. https://doi.org/10.1002/nano.202300038

[2] Wu Q, Miao WS, Zhang YD, Gao HJ, Hui D. Mechanical properties of nanomaterials: A review. Nanotechnology Reviews. 2020 Mar 24;9(1):259-73. https://doi.org/10.1515/ntrev-2020-0021

[3] Amodeo J, Pizzagalli L. Modeling the mechanical properties of nanoparticles: a review. Comptes Rendus. Physique. 2021;22(S3):35-66. https://doi.org/10.5802/crphys.70

[4] Awan MS, Anwar MT, Khan HI, Asghar MR, Raza MR, Husnain N, Hussain M, Rasheed T. Nanoparticles' classification, synthesis, characterization and applications—a review. Charact. Appl. Nanomater. 2024;8:8899. https://doi.org/10.24294/can8899

[5] Roy A, Gupta A, Haque B, Qureshi AA, Verma D, Sharma K, Lee SF, Hee CW, Roy A, Verma R. An updated review on carbon nanomaterials: Types, synthesis, functionalization and applications, degradation and toxicity. Green Processing and Synthesis. 2024 Oct 29;13(1):20240150. https://doi.org/10.1515/gps-2024-0150

[6] Vogl LM, Schweizer P, Richter G, Spiecker E. Effect of size and shape on the elastic modulus of metal nanowires. MRS Advances. 2021 Sep;6(27):665-73. https://doi.org/10.1557/s43580-021-00103-3

[7] Alasfar RH, Ahzi S, Barth N, Kochkodan V, Khraisheh M, Koç M. A review on the modeling of the elastic modulus and yield stress of polymers and polymer nanocomposites: Effect of temperature, loading rate and porosity. Polymers. 2022 Jan 18;14(3):360. https://doi.org/10.3390/polym14030360

[8] Eltaher MA, Khater ME, Emam SA. A review on nonlocal elastic models for bending, buckling, vibrations, and wave propagation of nanoscale beams. Applied Mathematical Modelling. 2016 Mar 1;40(5-6):4109-28. https://doi.org/10.1016/j.apm.2015.11.026

[9] Barretta R, Marotti de Sciarra F, Vaccaro MS. Nonlocal elasticity for nanostructures: a review of recent achievements. Encyclopedia. 2023 Feb 27;3(1):279-310. https://doi.org/10.3390/encyclopedia3010018

[10] Boyina K, Piska R. Wave propagation analysis in viscoelastic Timoshenko nanobeams under surface and magnetic field effects based on nonlocal strain gradient theory. Applied Mathematics and Computation. 2023 Feb 15;439:127580. https://doi.org/10.1016/j.amc.2022.127580

[11] He D, Shi D, Wang Q, Ma C. Wave propagation in magneto-electro-thermo-elastic nanobeams based on nonlocal theory. Journal of the Brazilian Society of Mechanical Sciences and Engineering. 2020 Nov;42(11):601. https://doi.org/10.1007/s40430-020-02683-x

Journal of Nanotechnology and Innovative Materials

Research Article

A Purely Dissipative Third-Order Exceptional Point in a Plasmonic Nanocavity with Two Quantum Dots: Exact Conditions and the Absence of an Enhanced Observable Response

Narmin A. Azizli

Department of Mechanics and Mathematics, School of Advanced Technologies and Innovation Engineering, Western Caspian University, 17A, Ahmad Rajabli Street, III Parallel, AZ1072 Baku, Azerbaijan

Received: 27.06.2026 Accepted: 26.07.2026 Published: 31.07.2026

https://doi.org/10.54414/WSLM7544

Abstract

Third-order exceptional points in cavity systems have so far been obtained by introducing optical or mechanical gain, which restricts them to platforms where gain can be engineered. We show that two semiconductor quantum dots in a lossy plasmonic nanocavity host such a point with loss alone, and we solve the coalescence problem for the 3 × 3 effective Hamiltonian in closed form. The conditions leave no free parameter: the cavity sits at the mean transition frequency of the dots, the direct dipole-dipole coupling vanishes, the loss contrast obeys $\kappa - \gamma = 3\sqrt{3}g$, and the two dots are detuned from each other by exactly one coupling constant, $\omega_1 - \omega_2 = g$. The tuning knob is the detuning between the emitters rather than a gain rate. We confirm that the effective Hamiltonian becomes a single Jordan block of size three and that the coalescence also appears in the spectrum of the full Liouvillian, where three eigenvalues agree to one part in 10^5 of the linewidth, and the geometric multiplicity remains one, so quantum jumps do not remove it. The eigenvalue splitting follows $(\varepsilon g^2)^{(1/3)}$ for a cavity perturbation ε, and measured against the linewidth it exceeds the second-order response of the same device by a factor of 4 to 12 over four decades. The steady-state cavity population, however, responds linearly in ε with a fitted exponent of 1.000 and an advantage over the second-order point of only 1.09. The reason is structural: the characteristic polynomial depends on ε linearly at fixed frequency, so the resolvent stays analytic even though the pole positions do not. The cube-root law therefore describes where the poles sit and not what a spectroscopic measurement returns. We also find that a residual dipole-dipole coupling of 10^{-3} g restores linear eigenvalue response, which sets a minimum interdot separation of 12 to 36 nm and conflicts with the tight confinement that produces large coupling in plasmonic gaps.

Keywords: non-Hermitian physics, third-order exceptional point, Liouvillian spectrum, coupled quantum dots, plasmonic nanocavity, Lindblad master equation

1. Introduction

An exceptional point is a parameter value at which several eigenvalues of a non-Hermitian operator become equal, and their eigenvectors coalesce simultaneously [1]. The operator is then defective, and the spectral response to a perturbation ε loses its analyticity: for a coalescence of order n, the eigenvalues separate as $\varepsilon^{(1/n)}$. Optical and photonic systems have been the main testing ground, and the general theory is reviewed by Miri and Alù [2] and by Özdemir and co-workers [3].

Third-order points attract attention because the exponent 1/3 promises a larger response than the square root of a second-order point. They were first reached in coupled microring resonators with balanced gain and loss [4], and third-order exceptional points have also been investigated in layered optical microdisk cavities [5]. Proposals for quantum platforms followed. Almost all of them share a requirement. The pseudo-Hermitian condition used by Xiong and co-workers for a three-mode optomechanical system needs one resonator with gain [6], [7], the superconducting-circuit realisation of an exceptional surface uses balanced gain and loss [8], and the ion-cavity scheme of Kim and co-workers relies on an engineered pumping configuration [9]. Gain is not available in every platform, and in plasmonic nanostructures it is particularly awkward.

This paper starts from a system in which there is no gain at all. Two semiconductor quantum dots sit in a single-mode plasmonic nanocavity, each coupled to the cavity with strength g, with a possible direct dipole-dipole term J between them. Room-temperature strong coupling between single emitters and plasmonic gap modes is established experimentally [10], [11], with couplings of tens of meV against cavity linewidths of a hundred meV or more, and the broader physics of strong coupling between surface plasmon polaritons and quantum emitters has been reviewed extensively [12]. In the single-excitation sector, this is a three-level non-Hermitian problem, the smallest that can support a third-order coalescence, and every diagonal imaginary part is a loss.

We answer two questions. The first is whether the coalescence exists at all in such a system and under what conditions, and the answer is a single isolated point with a closed form in which the detuning between the two dots plays the role that a gain rate plays elsewhere. The second question is what the coalescence is worth. Here our result is negative, and we consider it the more useful of the two. The cube-root law governs the positions of the poles, but the steady-state optical response of the device is linear in ε, with essentially no advantage over a second-order point. This is a concrete instance of the general warning raised by Langbein [13] and by Lau and Clerk [14], worked out for a specific system rather than in the abstract.

2. Model

2.1. Hamiltonian and Master Equation

Two quantum dots with transition frequencies ω_1 and ω_2 couple to a cavity mode of frequency ω_c. In the rotating wave approximation, and with ħ = 1, the Hamiltonian in a frame rotating at the drive frequency ω_L reads

$$H = \Delta_c a^\dagger a + \Delta_1 \sigma_1^+ \sigma_1^- + \Delta_2 \sigma_2^+ \sigma_2^- + g\left(a^\dagger \sigma_1^- + a\sigma_1^+\right) + g\left(a^\dagger \sigma_2^- + a\sigma_2^+\right) + J(\sigma_1^+ \sigma_2^- + \sigma_2^+ \sigma_1^-) + E\left(a + a^\dagger\right) \quad (1)$$

where $\Delta_c = \omega_c - \omega_L$ and $\Delta_i = \omega_i - \omega_L$ are the detunings, a is the cavity annihilation operator, $\sigma_i^\pm$ are the raising and lowering operators of dot i, J is the direct dipole-dipole coupling, and E is the amplitude of a weak coherent drive. Dissipation is treated with the Markovian master equation [15]

$$\frac{d\rho}{dt} = -i[H,\rho] + \kappa\mathcal{D}[a]\rho + \gamma\mathcal{D}[\sigma_1^-]\rho + \gamma\mathcal{D}[\sigma_2^-]\rho \quad (2)$$

with the dissipator

$$\mathcal{D}[A]\rho = A\rho A^\dagger - \frac{1}{2}\left\{A^\dagger A,\rho\right\} \quad (3)$$

Here κ is the cavity photon loss, dominated by ohmic absorption in the metal, and γ is the spontaneous emission rate of each dot into modes other than the cavity. Both rates are positive. No gain enters anywhere.

2.2. Effective Non-Hermitian Hamiltonian

Set E = 0 and restrict the dynamics to the single-excitation subspace spanned by |1, g, g⟩, |0, e, g⟩ and |0, g, e⟩. Between quantum jumps, the state evolves under

$$H_{eff} = \begin{bmatrix} \omega_c - i\frac{\kappa}{2} & g & g \\ g & \omega_1 - i\frac{\gamma}{2} & J \\ g & J & \omega_2 - i\frac{\gamma}{2} \end{bmatrix} \quad (4)$$

Introducing

$$\omega_{1,2} = \omega_0 \pm \delta, \quad A = \Delta_c - iq, \quad q = \frac{\kappa - \gamma}{2}, \quad \Delta_c = \omega_c - \omega_0 \tag{5}$$

and subtracting $\omega_0 - i\gamma/2$ from the diagonal, the shifted matrix becomes

$$H' = \begin{bmatrix} A & g & g \\ g & \delta & J \\ g & J & -\delta \end{bmatrix} \tag{6}$$

The whole of the non-Hermiticity now sits in the single complex number A while every other entry is real, which is what makes the problem solvable by hand.

3. Exact Conditions for the Coalescence

The characteristic polynomial of H′ is $\lambda^3 + c_2\lambda^2 + c_1\lambda + c_0$ with

$$c_2 = -A, \quad c_1 = -(2g^2 + \delta^2 + J^2), \quad c_0 = A(\delta^2 + J^2) - 2g^2 J \tag{7}$$

A triple root requires the polynomial to equal $(\lambda - \lambda_0)^3$ with $\lambda_0 = -c_2/3 = A/3$, which gives the two complex conditions $c_1 = c_2^2/3$ and $c_0 = c_2^3/27$. The first reads

$$A^2 = -3S, \quad S = 2g^2 + \delta^2 + J^2 \tag{8}$$

S is real and positive, so A must be purely imaginary. The cavity therefore has to be tuned to the mean transition frequency of the dots, $\Delta_c = 0$, and writing $A = -iq$, the condition becomes $q^2 = 3S$. Substituting into the second condition and separating real from imaginary parts gives

$$2g^2 J = 0, \qquad \delta^2 + J^2 = \frac{q^2}{27} \tag{9}$$

The first of these forces J = 0, which is the least expected feature of the result. The coalescence survives only when the two dots have no direct coupling to each other, and any residual dipole-dipole interaction removes it. With J = 0 the remaining relations close and give $q^2 = 27g^2/4$ and $\delta^2 = g^2/4$, so the third-order point occurs at

$$\Delta_c = 0, \quad J = 0, \quad \kappa - \gamma = 3\sqrt{3}\, g, \quad \omega_1 - \omega_2 = 2\delta = g \tag{10}$$

with the threefold eigenvalue

$$\lambda_0 = \omega_0 - i\left(\frac{\gamma}{2} + \frac{\sqrt{3}\, g}{2}\right) \tag{11}$$

Every parameter is fixed once g is chosen, and the only remaining freedom is the overall energy scale. Two features are worth stating plainly. The loss contrast has to be about 5.2 times the coupling strength, which is a regime plasmonic cavities reach naturally. The two dots have to be detuned by exactly one coupling constant, so the emitters are neither degenerate nor far apart, and this detuning is the parameter that replaces the gain rate of earlier proposals [6], [7], [8], [9].

We verified equation (10) numerically for g = 30 meV. The three eigenvalues agree with λ_0 to 3×10^{-4} meV, which is the accuracy any eigenvalue routine can reach near a defective matrix since the error scales as the cube root of the machine epsilon. The matrix $H' - \lambda_0 I$ has rank two, so only one eigenvector survives, and $(H' - \lambda_0 I)^3$ vanishes to 1.9×10^{-11} in the Frobenius norm. The Jordan form is a single block of size three. Figure 1 sweeps the dot-dot detuning at fixed loss contrast, and the real and imaginary parts of all three branches meet at $\delta = g/2$.

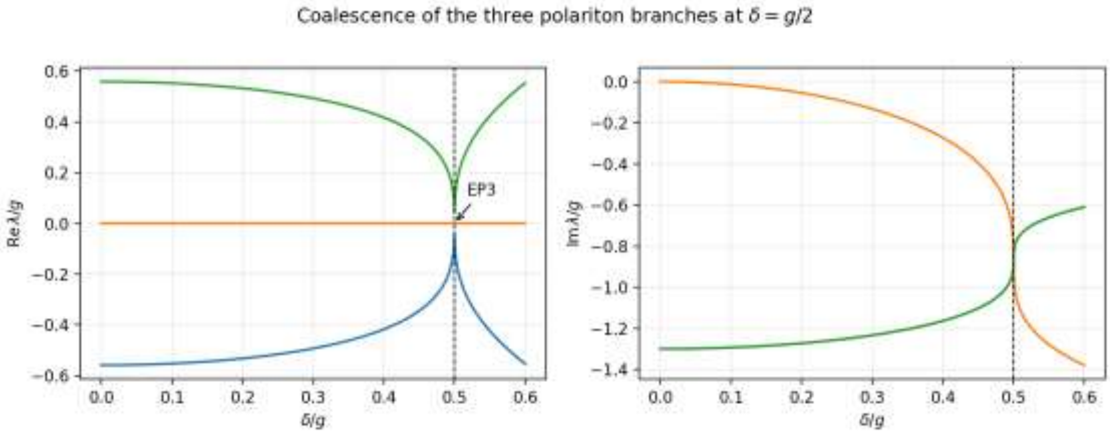

Figure 1. Real and imaginary parts of the three eigenvalues of H′ against the dot-dot detuning δ, at $\Delta\_c = 0$, J = 0, and $\kappa - \gamma = 3\sqrt{3}g$. The three branches meet at $\delta = g/2$.

Figure 2 maps the discriminant of the characteristic polynomial over the plane of δ and loss contrast. The dark curves are second-order exceptional lines, and they meet in a cusp at the point of equation (10).

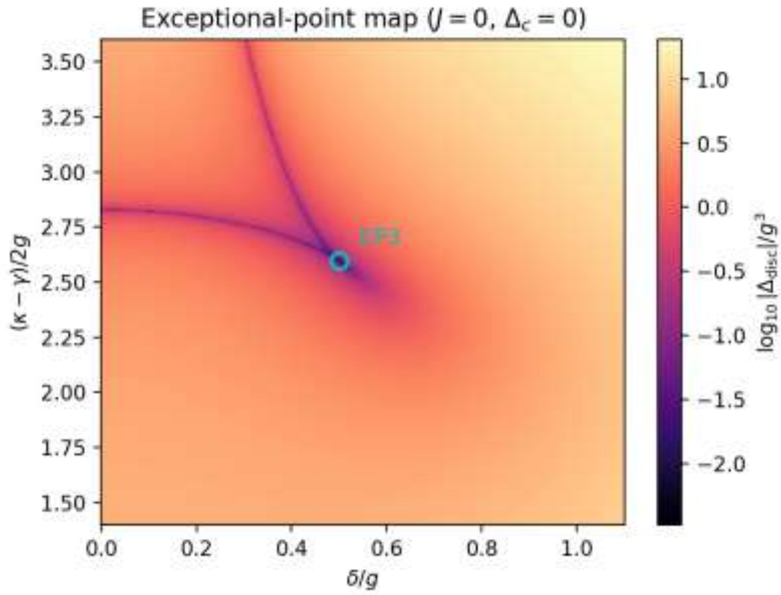

Figure 2. Logarithm of the discriminant of the characteristic polynomial in the $(\delta, \kappa - \gamma)$ plane for J = 0 and $\Delta\_c = 0$. The two second-order lines meet at the third-order point marked in cyan.

4. The Coalescence in the Full Liouvillian

The effective Hamiltonian of equation (4) drops the recycling term of the master equation, so a coalescence found there need not survive in the full dynamics. Exceptional points of a Liouvillian and of the corresponding non-Hermitian Hamiltonian are known to differ in general. We therefore diagonalised the Liouvillian of equation (2) directly, on a Hilbert space of dimension 3 × 2 × 2 with the cavity truncated at three photons.

The single-excitation coherences of the density matrix appear in the Liouvillian spectrum at $\mu = -i\lambda$, where λ are the eigenvalues of H_eff. At the parameters of equation (10), the three Liouvillian eigenvalues closest to that prediction lie within 5.6×10^{-4} meV of each other, which is 1.1×10^{-5} of the coalescence linewidth. The matrix $\mathscr{L} - \mu I$ has only one vanishing singular value at that eigenvalue, so its geometric multiplicity is one, and the Liouvillian is defective there as well. Quantum jumps shift the coalescence by an amount far below any linewidth and do not destroy it. This is the first of the two checks that a reader would reasonably demand.

5. Eigenvalue Response and its Resolvability

An analyte binding to the metal surface changes the local permittivity by δε_r and shifts the cavity frequency by

$$\varepsilon = \frac{\omega_c\, \delta\varepsilon_r}{2{n_0}^2} \qquad (12)$$

which perturbs the (1,1) element of H′ alone. Expanding the characteristic polynomial about the coalescence with the cofactor of that element gives

$$(\lambda - \lambda_0)^3 = \varepsilon\left({\lambda_0}^2 - \delta^2\right) = -\varepsilon g^2 \qquad (13)$$

since $\lambda_0 = -i\sqrt{3}g/2$ and $\delta = g/2$ in the shifted frame. The roots therefore separate as

$$|\lambda - \lambda_0| = (\varepsilon g^2)^{\frac{1}{3}}, \quad |\Delta\lambda_{max}| = \sqrt{3}(\varepsilon g^2)^{\frac{1}{3}} \qquad (14)$$

with the three branches spaced by 120 degrees in the complex plane. Figure 3 compares this with the second-order point of the same device, reached by making the dots degenerate so that the dark state decouples and the cavity couples to the bright state with $\sqrt{2}g$ [16]. That configuration has its exceptional point at $\kappa - \gamma = 4\sqrt{2}g$ and responds as the square root of ε. Table 1 lists the numbers.

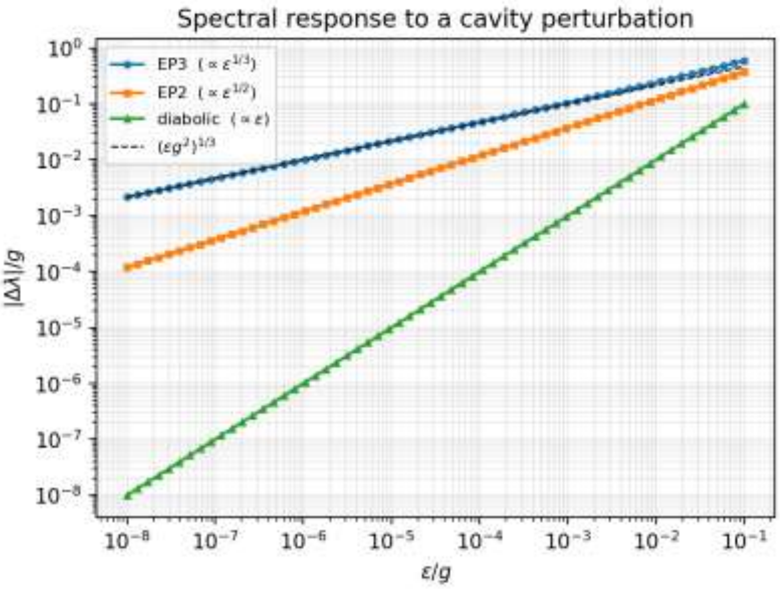

Figure 3. Eigenvalue splitting against the cavity perturbation ε on logarithmic axes. The slopes are 1/3, 1/2, and 1. The dashed line is equation (14).

Table 1. Eigenvalue splitting for the three cases of Figure 3.

ε / g	EP3, |δλ| / g	EP2, |δλ| / g	Conventional, |δω| / g	EP3 / EP2
10^{-2}	2.409×10^{-1}	1.189×10^{-1}	10^{-2}	2.03
10^{-3}	1.052×10^{-1}	3.761×10^{-2}	10^{-3}	2.80
10^{-4}	4.751×10^{-2}	1.189×10^{-2}	10^{-4}	4.00
10^{-5}	2.178×10^{-2}	3.761×10^{-3}	10^{-5}	5.79
10^{-6}	1.005×10^{-2}	1.189×10^{-3}	10^{-6}	8.45

A splitting is only meaningful next to a linewidth, and the three cases have very different linewidths. At the third-order point the coalesced triple has a full width of $\sqrt{3}g$, at the second-order point the width is $2\sqrt{2}g$, and a conventional measurement on the same platform works against the bare cavity width κ. Normalising each splitting by its own width gives

$$R_3 = \left(\frac{\varepsilon}{g}\right)^{\frac{1}{3}}, \qquad R_2 = 0.841 \left(\frac{\varepsilon}{g}\right)^{\frac{1}{2}} \tag{15}$$

So the advantage of the third-order point survives the normalisation and in fact grows, because its linewidth is the narrowest of the three. Table 2 and Figure 4 give the numbers. Against the second-order point of the same device, the gain runs from 2.8 at $\varepsilon = 10^{-2}$ g to 12 at $\varepsilon = 10^{-6}$ g. Against a conventional shift measurement in the same cavity, it is five orders of magnitude. The splitting reaches ten percent of the linewidth at $\varepsilon = 0.026$ meV for g = 30 meV, which is a spectroscopically plausible target.

Table 2. Splitting normalised by the corresponding linewidth.

ε / g	EP3, split / FWHM	EP2, split / FWHM	Conventional, shift / FWHM	EP3 / EP2
10^{-2}	2.390×10^{-1}	8.409×10^{-2}	1.924×10^{-3}	2.84
10^{-3}	1.050×10^{-1}	2.659×10^{-2}	1.924×10^{-4}	3.95
10^{-4}	4.750×10^{-2}	8.409×10^{-3}	1.924×10^{-5}	5.65
10^{-5}	2.178×10^{-2}	2.659×10^{-3}	1.924×10^{-6}	8.19
10^{-6}	1.005×10^{-2}	8.409×10^{-4}	1.924×10^{-7}	11.95

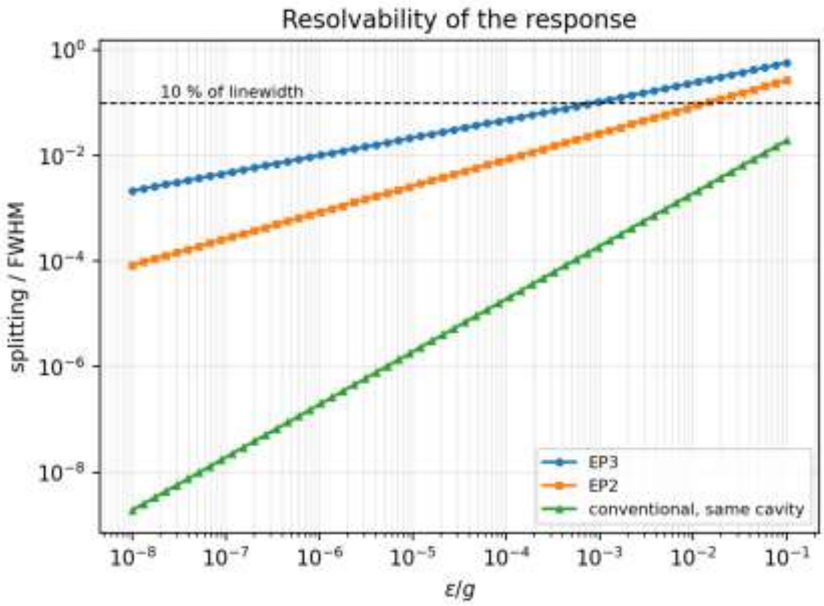

Figure 4. Splitting divided by the full width at half maximum of the relevant resonance, for the third-order point, the second-order point, and a conventional shift measurement in the same cavity.

6. Response of a Measurable Quantity

Everything in section 5 concerns eigenvalues, which are not what a spectrometer records. To find the response of an observable, we solved the master equation for the steady state under a weak coherent drive, $\mathscr{L}\rho$_ss = 0, by direct linear inversion with the trace condition substituted for one row. The mean cavity population changes by less than 2×10^{-5} in relative terms between Fock cutoffs of three and five, so the truncation is converged at the drive amplitude used, E = 0.02 g.

Figure 5 shows the steady-state cavity population against drive detuning for three values of the cavity loss. Below the exceptional value, the spectrum has three resolved polariton peaks. At κ = κ_EP they merge into a single line, and above it the response is one broad peak.

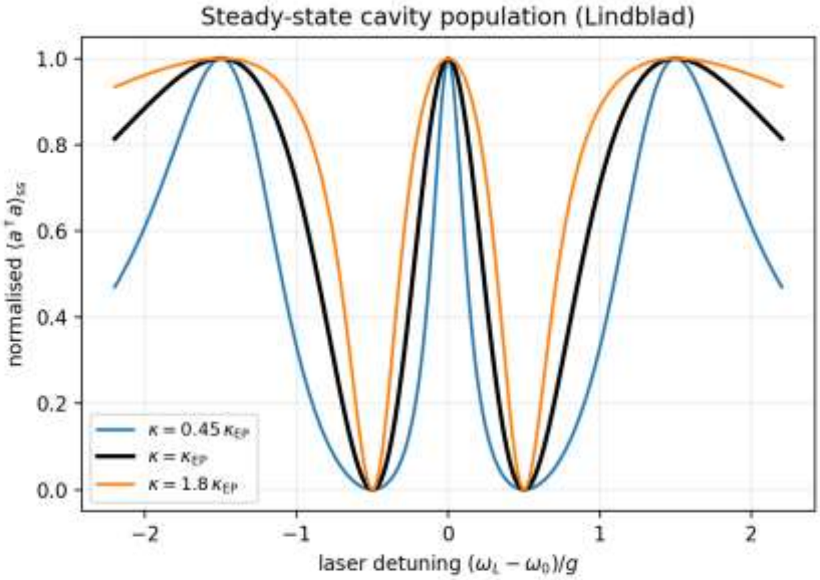

Figure 5. Normalised steady-state cavity population against laser detuning for three cavity loss rates.

We then applied the perturbation ε and measured how much the spectrum changes, taking the largest absolute change in the population normalised by its peak value. The result is Figure 6 and Table 3. Over more than two decades, the response follows a power law with a fitted exponent of 1.000, both at the third-order point and at the second-order point, and the ratio between the two is 1.09 and independent of ε. The cube-root behaviour is absent from the observable.

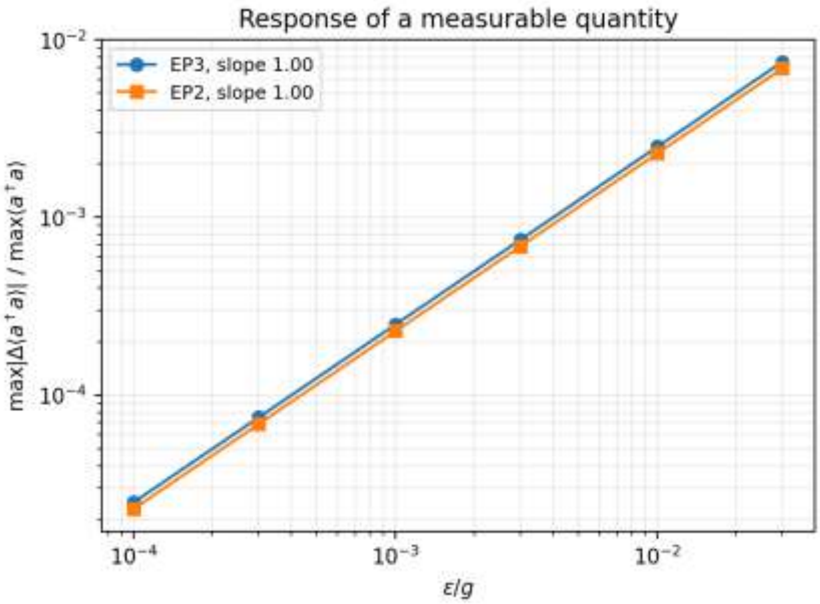

Figure 6. Relative change of the steady-state cavity population under a perturbation ε, at the third-order point and at the second-order point. Both are linear in ε.

Table 3. Response of the steady-state cavity population.

ε / g	EP3, max|Δ⟨n⟩| / ⟨n⟩max	EP2, max|Δ⟨n⟩| / ⟨n⟩max	EP3 / EP2
10^{-4}	2.498×10^{-5}	2.296×10^{-5}	1.09
10^{-3}	2.498×10^{-4}	2.296×10^{-4}	1.09
10^{-2}	2.499×10^{-3}	2.296×10^{-3}	1.09

The reason is structural rather than numerical, and it can be read off the characteristic polynomial. With the perturbation included,

$$P(\omega) = (\omega - \lambda_0)^3 - \varepsilon(\omega^2 - \delta^2), \qquad \frac{\partial P(\omega)}{\partial \varepsilon} = -(\omega^2 - \delta^2) \tag{16}$$

At a fixed observation frequency ω, the polynomial depends on ε linearly, and every steady-state response of the driven system is built from the resolvent, whose denominator is P(ω). The lineshape therefore changes at first order in ε. The cube root enters only when one asks where the roots of P have moved, and at a third-order point those three roots sit inside a single linewidth with residues that very nearly cancel, so no spectroscopic measurement separates them. Individual eigenvalues are not observables of the driven steady state.

This does not contradict section 5. It says that the normalised splitting of Figure 4 is the right figure of merit only for an experiment that can resolve the poles individually, for instance a time-domain measurement of the ringdown or a scheme that isolates one eigenvector. For the ordinary transmission or scattering measurement that an exceptional-point sensor is usually imagined to perform, the enhancement is not there. Langbein [13] and Lau and Clerk [14] reached the same conclusion for second-order points from noise arguments, and Wiersig [17] reviews the cases where an enhancement does survive. What the present calculation adds is that the obstruction appears already at the level of the deterministic response, before any noise is introduced.

7. Sensitivity to a Residual Dipole-Dipole

Because equation (9) forces J to vanish exactly, the tolerance on J is a practical question. A finite J splits the triple root into a second-order point plus a separate eigenvalue, and the residual splitting grows as the cube root of J, so it already reaches five percent of g at $J \approx 1.1 \times 10^{-3}$ g. At that value, the local response exponent measured at $\varepsilon = 10^{-5}$ g has returned to 0.99. Figure 7 shows both curves.

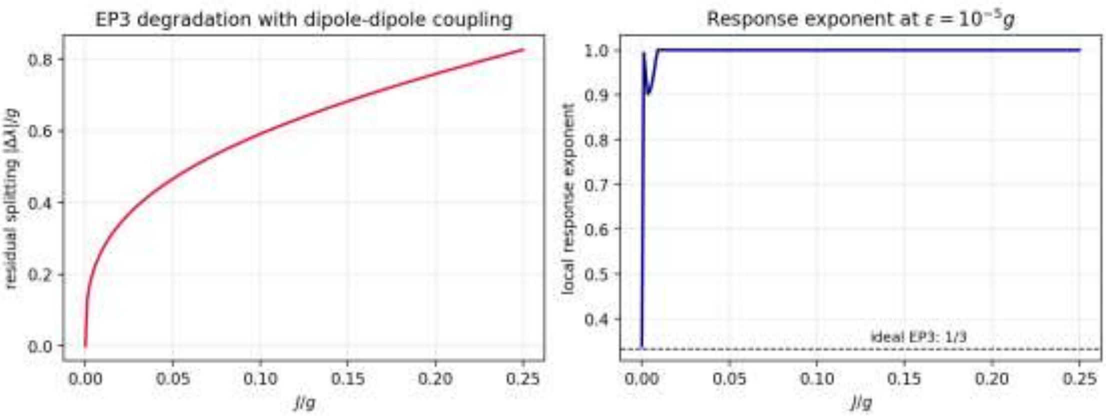

Figure 7. Left: residual eigenvalue splitting at the nominal exceptional point against the dipole-dipole coupling J. Right: local response exponent at $\varepsilon = 10^{-5}$ g, with the ideal value 1/3 marked.

For g = 30 meV, the tolerance is $|J| \lesssim 0.03$ meV. Treating the interaction as a point dipole coupling,

$$J = \frac{\mu^2}{4\pi\varepsilon_0 r^3} \tag{17}$$

this gives a minimum separation of about 12 nm for a transition dipole of 10 D, 20 nm for 20 D, and 26 nm for 30 D. The requirement pulls against the geometry that produces a large g, because plasmonic gap modes concentrate their field over a few nanometres. A device of this kind needs a cavity whose field stays strong across a few tens of nanometres, such as an extended nanoparticle-on-mirror gap or a bowtie antenna with a wider feed, and it pays for the separation with a smaller coupling. This tension sets the design window more tightly than any of the spectral conditions.

8. Parameter Estimate

Take $\omega_0 = 1.5$ eV and g = 30 meV, within the range reported for single emitters in plasmonic gap modes at room temperature [10], [11]. Equation (10) then requires κ = 155.9 meV, a quality factor of about 9.6, which is ordinary for a gold nanoparticle-on-mirror cavity. The dot-dot detuning must be 30 meV, comfortably inside the inhomogeneous distribution of self-assembled dots and reachable with a quantum-confined Stark shift on one dot. With $\gamma = 5$ μeV, the condition is effectively $\kappa = 3\sqrt{3}g$. The linewidth at the coalescence is $\sqrt{3}g/2 \approx 26$ meV from equation (11). None of these numbers is unusual, and the demanding parts are the separation constraint of section 7 and holding the detuning at g.

9. Conclusion

A two-dot plasmonic cavity supports a third-order exceptional point without any gain, and the conditions for it are simple enough to write down exactly: the cavity on resonance with the mean dot frequency, no direct dipole-dipole coupling, $\kappa - \gamma = 3\sqrt{3}g$, and a dot-dot detuning of exactly g. The detuning between the emitters takes the role that a gain rate plays in the optomechanical and circuit realisations reported so far. The coalescence is not an artefact of the effective non-Hermitian description, since the full Liouvillian is defective at the same parameters to within one part in 10^5 of the linewidth.

The eigenvalue splitting follows the expected cube-root law, and normalised against the linewidth it beats the second-order point of the same device by a factor of 4 to 12. The steady-state optical response does not follow it. That response is linear in the perturbation with an advantage of 1.09 over the second-order point, because the characteristic polynomial enters the resolvent linearly in ε at fixed frequency while the cube root lives only in the pole positions, which are unresolvable inside one linewidth. Any proposal that treats the ε^(1/3) eigenvalue law as a sensing figure of merit for a driven cavity measurement should be read with this in mind.

Two directions follow. A time-domain measurement is not bound by the argument of section 6 and might recover the cube root, which would be worth calculating. The tolerance on the dipole-dipole coupling, one part in a thousand of g, is the practical obstacle, and a cavity geometry that keeps a large coupling across a few tens of nanometres would be the first thing such an experiment needs.

Appendix. Numerical methods

All calculations use numpy and scipy. Eigenvalues of the 3 × 3 effective Hamiltonian come from numpy.linalg.eigvals. Near a defective matrix, the accuracy of any eigenvalue routine scales as the cube root of the machine epsilon, which explains the 10^{-4} meV residual quoted in section 3, so the Jordan structure is confirmed instead through the rank of $H' - \lambda_0 I$ and the norm of its cube. The Liouvillian is assembled by column-stacking vectorisation,

$$\mathcal{L} = -i(H \otimes \mathbb{1} - \mathbb{1} \otimes H^T) + \Sigma_k \left[C_k \otimes C_k^* - \tfrac{1}{2}C_k^\dagger C_k \otimes \mathbb{1} - \tfrac{1}{2}\mathbb{1} \otimes \left(C_k^\dagger C_k\right)^T\right] \tag{18}$$

where the collapse operators are $\sqrt{\kappa}\, a$, $\sqrt{\gamma}\, \sigma_1^-$ and $\sqrt{\gamma}\, \sigma_2^-$. The steady state follows from replacing the first row of $\mathscr{L}$ with the vectorised identity and solving against a unit vector. Liouvillian eigenvalues come from dense diagonalisation of the 144 × 144 matrix and its defectiveness from the singular values of $\mathscr{L} - \mu I$. The scripts are supplied as supplementary material and reproduce every figure and number in the text.

Author Contributions

The author contributed to the conception, design, analysis, interpretation, and preparation of the manuscript and approved the final version.

Funding

No funding was received for this research.

Conflicts of Interest

The author declares no conflict of interest.

Acknowledgment

Not applicable.

Abbreviations

Exceptional Point (EP), Second-Order Exceptional Point (EP2), Third-Order Exceptional Point (EP3), Full Width at Half Maximum (FWHM), Half Width at Half Maximum (HWHM), Quantum Dot (QD), Rotating-Wave Approximation (RWA).

References

[1] Heiss WD. The physics of exceptional points. Journal of Physics A: Mathematical and Theoretical. 2012 Nov 9;45(44):444016. https://doi.org/10.1088/1751-8113/45/44/444016
[2] Miri MA, Alu A. Exceptional points in optics and photonics. Science. 2019 Jan 4;363(6422):eaar7709. https://doi.org/10.1126/science.aar7709
[3] Özdemir ŞK, Rotter S, Nori F, Yang L. Parity–time symmetry and exceptional points in photonics. Nature Materials. 2019 Aug;18(8):783-98. https://doi.org/10.1038/s41563-019-0304-9
[4] Hodaei H, Hassan AU, Wittek S, Garcia-Gracia H, El-Ganainy R, Christodoulides DN, Khajavikhan M. Enhanced sensitivity at higher-order exceptional points. Nature. 2017 Aug 10;548(7666):187-91. https://doi.org/10.1038/nature23280
[5] Kullig J, Yi CH, Hentschel M, Wiersig J. Exceptional points of third-order in a layered optical microdisk cavity. New Journal of Physics. 2018 Aug 1;20(8):083016. https://doi.org/10.1088/1367-2630/aad594
[6] Xiong W, Li Z, Song Y, Chen J, Zhang GQ, Wang M. Higher-order exceptional point in a pseudo-Hermitian cavity optomechanical system. Physical Review A. 2021 Dec;104(6):063508. https://doi.org/10.1103/physreva.104.063508
[7] Xiong W, Li Z, Zhang GQ, Wang M, Li HC, Luo XQ, Chen J. Higher-order exceptional point in a blue-detuned non-Hermitian cavity optomechanical system. Physical Review A. 2022 Sep;106(3):033518. https://doi.org/10.1103/physreva.106.033518
[8] Zhang G, Lin S, Feng W, Wang Y, Yu Y, Yang C. Third-order exceptional surface in a pseudo-Hermitian superconducting circuit. Science China Information Sciences. 2025 Aug;68(8):180508. https://doi.org/10.1007/s11432-025-4470-4
[9] Kim J, Ha T, Kim D, Lee D, Lee KS, Won J, Moon Y, Lee M. Third-order exceptional point in an ion–cavity system. Applied Physics Letters. 2023 Oct 16;123(16). https://doi.org/10.1063/5.0168372
[10] Chikkaraddy R, De Nijs B, Benz F, Barrow SJ, Scherman OA, Rosta E, Demetriadou A, Fox P, Hess O, Baumberg JJ. Single-molecule strong coupling at room temperature in plasmonic nanocavities. Nature. 2016 Jul 7;535(7610):127-30. https://doi.org/10.1038/nature17974
[11] Santhosh K, Bitton O, Chuntonov L, Haran G. Vacuum Rabi splitting in a plasmonic cavity at the single quantum emitter limit. Nature Communications. 2016 Jun 13;7(1):ncomms11823. https://doi.org/10.1038/ncomms11823
[12] Törmä P, Barnes WL. Strong coupling between surface plasmon polaritons and emitters: a review. Reports on Progress in Physics. 2015 Jan 1;78(1):013901. https://doi.org/10.1088/0034-4885/78/1/013901
[13] Langbein W. No exceptional precision of exceptional-point sensors. Physical Review A. 2018 Aug;98(2):023805. https://doi.org/10.1103/physreva.98.023805
[14] Lau HK, Clerk AA. Fundamental limits and non-reciprocal approaches in non-Hermitian quantum sensing. Nature Communications. 2018 Oct 17;9(1):4320. https://doi.org/10.1038/s41467-018-06477-7
[15] Breuer HP, Petruccione F. The theory of open quantum systems. OUP Oxford; 2002.
[16] Wiersig J. Enhancing the sensitivity of frequency and energy splitting detection by using exceptional points: application to microcavity sensors for single-particle detection. Physical Review Letters. 2014 May 23;112(20):203901. https://doi.org/10.1103/physrevlett.112.203901
[17] Wiersig J. Review of exceptional point-based sensors. Photonics Research. 2020 Sep 1;8(9):1457-67. https://doi.org/10.1364/prj.396115

Journal of Nanotechnology and Innovative Materials

Research Article

Investigation of the Effect of Rare Earth Elements in Hall Sensors Created Based on InSb

Ilham N. Salimov[1], Fatima H. Amiraslanli[1], and Zarifa I. Hasanli[1]

[1]Department of Mechanics and Mathematics, School of Advanced Technologies and Innovation Engineering, Western Caspian University, Ahmad Rajabli 17A, III Parallel, AZ1001 Baku, Azerbaijan

Received: 30.06.2026 Accepted: 22.07.2026 Published: 31.07.2026

https://doi.org/10.54414/LEGX7838

Abstract

The article investigates the effect of rare earth elements on the electrophysical properties of Hall sensors based on indium antimonide (InSb). The InSb semiconductor is widely used in modern sensor technologies due to its high electron mobility, small forbidden energy band gap, and high magnetic sensitivity. These properties make InSb one of the promising materials for the production of high-precision Hall sensors. During the study, the electrical conductivity, Hall coefficient, carrier concentration, and mobility of InSb crystals doped with rare earth elements such as samarium (Sm) and dysprosium (Dy) were studied over a wide temperature range. It was found that rare earth elements reduce the number of structural defects in the crystal lattice, optimize additional scattering centers, and increase the relaxation time of electrons. As a result, the mobility of electrons increases, the conductivity improves, and the sensitivity of Hall sensors increases. The experimental results were compared with theoretical models and found to be in high agreement. The studies have shown that the addition of samarium in an amount of about 0.1 atomic percent significantly improves the main technical parameters of the Hall sensor. This indicates that InSb-based rare-earth-doped semiconductors are promising materials for automotive electronics, aerospace engineering, medical devices, and industrial automation.

Keywords: InSb, Hall effect, Hall sensor, rare earth elements, samarium, dysprosium, electron mobility, electrophysical properties

1. Introduction

The development of modern electronics requires the preparation of semiconductor materials with new functional capabilities and an in-depth study of their physical properties. The main part of the sensors used in microelectronics, automation, aerospace engineering, medical diagnostics, and industrial control systems is made up of highly sensitive semiconductor elements. In this regard, semiconductor compounds belonging to the A^3B^5 group are of particular interest [1].

Among these materials, indium antimonide (InSb) stands out due to its unique physical properties. InSb is one of the semiconductors with the highest electron mobility at room temperature [2]. The high electron mobility increases the sensitivity to the influence of a magnetic field, and this property is an important advantage for the development of sensors based on the Hall effect.

Another important advantage of InSb crystals is their narrow energy band gap. This property allows for high electrical conductivity with low power consumption. At the same time, it is possible to register even small magnetic fields [3] accurately. Therefore, Hall sensors made on the basis of InSb are widely used in measuring the rotational speed of cars, controlling electric motors, robotics, medical equipment, and aerospace systems.

In recent times, the application of various dopant elements has become widespread in order to improve the functional properties of semiconductors. In particular, rare earth elements have a significant effect on the electrical, magnetic, and optical properties of the material by changing the structure of the crystal lattice. Among these elements, samarium, dysprosium, erbium, gadolinium, and neodymium are of particular interest.

The difference in atomic radii and electronic structure of rare earth elements causes them to create new local energy levels in the crystal. As a result, the scattering mechanisms of charge carriers change, the relaxation time of electrons increases, and the mobility increases. This is of great scientific and practical importance in terms of increasing the sensitivity of Hall sensors [4].

The main objective of the present study is to study the electrophysical properties of InSb crystals doped with samarium and dysprosium, to determine the temperature dependence of the Hall coefficient, and to compare the obtained results with theoretical models.

InSb is a narrow-bandgap semiconductor belonging to the A^3B^5 type compound group. It has a zinc blende-type cubic crystal structure and is widely used in electronic engineering due to its high crystallinity.

The main advantage of InSb is the very high electron mobility. The electron mobility at room temperature is about 7.5×10^4–8.0×10^4 cm²/(V·s). High mobility leads to an increase in the Hall voltage in the magnetic field and an increase in the sensitivity of the sensor [5]. The band gap of InSb is about 0.17 eV. This small energy difference ensures the formation of charge carriers even at low temperatures. At the same time, a high sensitivity to infrared radiation is formed, which allows the InSb material to be widely used in infrared detectors. The effective mass of electrons in InSb crystals is very small (about 0.014 m_0) [6]. The small effective mass facilitates the acceleration of charge carriers and leads to an increase in electrical conductivity.

Electrical conductivity is defined by the following expression:

$$\sigma = q\left(n\mu_n + p\mu_p\right)$$

where:
σ – electrical conductivity;
q – electron charge;
n and p – electron and hole concentrations;
μ_n and μ_p – electron and hole mobilities.

The high magnetoresistive effect of InSb is also one of its important advantages. In a strong magnetic field, the trajectory of electrons is bent and the Hall voltage increases. Therefore, Hall sensors made on the basis of InSb allow for the measurement of even small magnetic fields with high accuracy.

One of the most effective methods for improving the physical properties of semiconductor materials is their modification with various additive elements. In particular, rare earth elements (Rare Earth Elements – REE) have been widely studied in recent years in terms of controlling the electrophysical and magnetic properties of narrow-bandgap semiconductors. Elements such as samarium (Sm), dysprosium (Dy), gadolinium (Gd), erbium (Er), and neodymium (Nd) create local energy levels in the crystal lattice due to their incompletely filled 4f-electron levels, which have a significant effect on the behavior of charge carriers. Adding small amounts (0.01–0.5 at.%) of rare earth elements to the InSb crystal optimizes its electrophysical parameters without causing significant changes in the structure of the crystal lattice.

The Hall effect was discovered in 1879 by the American physicist Edwin Hall and later became one of the fundamental phenomena in semiconductor physics. The Hall effect is characterized by the generation of an electric potential difference on the side surfaces of a sample when an electric current passes through a conductor or semiconductor and a magnetic field is applied perpendicular to it. This potential difference is called the Hall voltage and is defined by the following expression:

$$U_H = \frac{R_H I B}{d}$$

where:
U_H – Hall voltage;
R_H – Hall coefficient;
I – current flowing through the sample;
B – magnetic induction;
d – sample thickness.

The Hall coefficient is defined by the expression (for a material with one type of charge carrier).

$$R_H \approx \frac{1}{qn}$$

Here q is the elementary charge, and n is the concentration of charge carriers.

From this formula, it can be seen that the Hall coefficient is inversely proportional to the concentration of charge carriers. As the concentration of charge carriers decreases, the Hall coefficient increases and the sensitivity of the sensor increases [7].

The high electron mobility of the InSb semiconductor causes the Hall voltage to take larger values. Therefore, Hall sensors made on the basis of InSb allow for the measurement of weak magnetic fields with such high accuracy.

2. Materials and Methods

The samples studied in this work were prepared on the basis of high-purity InSb single crystals grown by the Bridgman method. Some of the crystals were synthesized with pure InSb, and the other part with samarium and dysprosium additives in various concentrations. The doping concentration was varied in the range of 0.01–0.30 atomic percent, and the measurements were carried out in the temperature range of 77–450 K. The electrical conductivity was measured by the four-probe method, and the Hall coefficient was measured in a constant magnetic field according to the classical Hall scheme.

The experimental results showed that when the amount of samarium additive is about 0.1 atomic percent, the temperature dependence of the Hall coefficient becomes more stable. The electron mobility increases significantly compared to the initial InSb crystals, and the electrical resistance decreases.

Similar results were observed in samples doped with dysprosium. However, the large magnetic moment of Dy atoms led to an even higher sensitivity to changes in the magnetic field. This feature is especially important for the development of high-precision magnetic sensors.

The obtained results showed that in the low temperature range (77–200 K) the scattering of electrons is mainly determined by ionized dopant atoms. As the temperature increases, the role of lattice phonons increases, and the electron-phonon interaction dominates. The addition of rare earth elements partially reduces the intensity of phonon scattering and increases the mean free path of electrons [8]. It was found that the mobility of electrons in samarium-doped InSb crystals increases several times compared to pure InSb in some temperature ranges. As a result, the Hall voltage increases, the sensitivity of the sensor increases, and the stability of the output signal improves.

Comparison of experimental data with theoretical calculations showed that the used model describes the change in the Hall coefficient depending on the temperature and dopant concentration with a sufficiently high accuracy [9]. This agreement confirms the correctness of the chosen physical model and creates an important scientific basis for future applications of InSb–Sm, as well as InSb–Dy systems.

3. Theoretical Explanation of the Results

The band theory of semiconductors is used to explain the effect of rare earth elements in InSb-based semiconductors. Since InSb has a narrow band gap ($E_g \approx 0.17$ eV), the concentration of charge carriers at room temperature is relatively high. When rare earth elements such as samarium (Sm) and dysprosium (Dy) are introduced into the crystal lattice, their incompletely filled 4f-electron levels create local energy states. These states affect the position of the Fermi level and cause a change in the concentration of charge carriers.

According to the Fermi–Dirac distribution law, the distribution of electrons over energy levels is described by the expression

$$f(E) = \frac{1}{1 + \exp[(E - E_F)/kT]}$$

where E_F is the Fermi level, k is the Boltzmann constant, and T is the absolute temperature.

The addition of rare earth elements shifts the Fermi level, which directly affects the concentration of electrons and holes. The movement of electrons within the crystal is mainly limited by two scattering mechanisms: ionized impurity atoms and lattice vibrations (phonons). At low temperatures, ionized impurity scattering dominates. As the temperature increases, the number of phonons increases, and the electron-phonon interaction becomes the main scattering mechanism. The introduction of rare earth elements into the crystal in small concentrations leads to compensation of structural defects. As a result, the relaxation time (τ) of the electrons increases. The electron mobility

$$\mu = \frac{q\tau}{m^*}$$

is directly proportional to the relaxation time and inversely proportional to the effective mass.

Here m* is the effective mass of the electron. An increase in the relaxation time increases the mobility of electrons, which leads to an improvement in the Hall coefficient and electrical conductivity. Electrical conductivity

$$\sigma = q\left(n\mu_n + p\mu_p\right)$$

is determined by the expression. Since the mobility of electrons in InSb is much higher than the mobility of holes, the conductivity is mainly determined by electrons. As a result of the influence of samarium and dysprosium additions, the mean free path of electrons increases, which leads to an increase in conductivity and an improvement in the sensitivity of the Hall sensor [10], [11], [12], [13]. Thus, theoretical calculations confirm the experimental results and show that doping with rare earth elements is one of the main factors optimizing the electrophysical properties of InSb crystals.

4. Discussion

Experimental studies have shown that InSb crystals doped with samarium and dysprosium have higher electron mobility, lower resistivity, and greater Hall sensitivity than pure InSb samples. The results obtained are generally consistent with the results obtained by other researchers on A^3B^5-type semiconductors. Various authors have shown that rare earth elements provide compensation for defects in the crystal lattice, increase the relaxation time of electrons, and reduce the scattering of charge carriers. The results of this study also confirm these regularities.

It has been established that when the doping concentration exceeds the optimal value, new defect centers are formed in the crystal and electron scattering increases. This leads to a decrease in the Hall coefficient and a deterioration in electron mobility. Therefore, optimizing the dopant concentration is considered an important technological issue in the preparation of high-quality Hall sensors. Samples doped with samarium at a level of about 0.1 atomic percent showed the best results. At this concentration, electron mobility increases, Hall voltage increases, and temperature stability improves. The addition of dysprosium, since it has a high magnetic moment, makes an additional contribution to increasing magnetic sensitivity.

The application areas of Hall sensors in the modern electronics industry are rapidly expanding. They are widely used in electric vehicles, robotics, industrial automation, medical diagnostic equipment, aerospace systems, and high-precision position sensors. In this regard, InSb crystals modified with rare earth elements can be considered a promising material for next-generation sensor technologies. In addition, in future studies, it would be appropriate to systematically investigate the effect of not only Sm and Dy, but also other rare earth elements such as Gd, Er, Nd, and Ho on the electrophysical and magnetic properties of InSb crystals.

5. Conclusion

As a result of the conducted studies, it was found that the electrophysical properties of InSb-based Hall sensors significantly depend on the degree of doping with rare earth elements. Samarium and dysprosium additions reduce the number of structural defects in the crystal lattice, increase the relaxation time of electrons, and weaken the scattering of charge carriers. As a result, electron mobility increases, electrical conductivity increases, and the sensitivity of the Hall sensor improves.

Comparison of experimental results with theoretical calculations showed that the selected physical model describes the dependence of the Hall coefficient and electron mobility on temperature with a sufficiently high accuracy. This agreement confirms the scientific basis of the conducted studies.

It was determined that an addition of about 0.1 atomic percent of samarium can be considered the optimal doping level for InSb crystals. At this concentration, the sensitivity, temperature stability, and reliability of the Hall sensor are close to the maximum values.

Thus, InSb-based semiconductors modified with rare earth elements are considered promising materials for the preparation of highly sensitive Hall sensors, and their wide application in the fields of modern electronics, automotive industry, space technology, and industrial automation is possible.

Author Contributions

All authors reviewed, edited, and approved the manuscript.

Conflict of Interest

The authors declare no conflicts of interest.

Funding

This research received no external funding.

Acknowledgment

The authors declare no acknowledgments.

Abbreviations

Indium Antimonide (InSb), Samarium (Sm), Dysprosium (Dy), Gadolinium (Dy), Erbium (Er), Neodymium (Nd), Rare-Earth Elements (REE).

References

[1] Askerov BM. Kinetic effects in semiconductors. Leningrad: Nauka; 1970. 303 p.
[2] Aliyev MI, Jabbarov RM. Electrical properties of GaSb and InSb doped with Gd and Yb eutectics based on them. In: Physical properties of complex semiconductors. Baku: Elm; 1982. p. XX–XX.
[3] Iskender-zade ZA, Salimov IN, Jabbarov RM, Gashim-zade FM. Preparation and electrophysical properties of samarium-doped indium antimonide. Inorganic Materials. 1991;27(2).
[4] Ashcroft NW. Solid State Physics. Thomson Learning. 1976;39.
[5] Kittel C. Solid State Physics, 8th edn. Hoboken, 2005.
[6] Ng KK, Sze SM. Physics of semiconductor devices. Hoboken, NJ: Wiley-Interscience; 2007.
[7] Look DC. Electrical characterization of GaAs materials and devices, 1989.
[8] Blakemore JS. Semiconductor statistics. Courier Corporation; 2002.
[9] Adachi S. Properties of semiconductor alloys: group-IV, III-V and II-VI semiconductors. John Wiley & Sons; 2009 Apr 27.
[10] Rogalski A. Infrared and terahertz detectors. CRC Press; 2019 Jan 10.
[11] Kasap SO. Electronic Materials and Devices. McGraw-Hill New York; 2006.
[12] Yu P. Fundamentals of semiconductors. USA; 2005.
[13] Ziman JM. Electrons and phonons, Oxford Classics Series, 2001.

Western Caspian University

Journal of Nanotechnology and Innovative Materials

Vol 2, No 1
July, 2026

Editorial Office Address: 17 A, Ahmad Rajabli Street,

III Parallel, Baku, Azerbaijan

Phone: +994 12 565 39 77

Email: jnim@wcu.edu.az

https://jnim.wcu.edu.az/

Submitted for collection: 23.06.2026

Signed for printing: 31.07.2026

Paper format: 60x84 1/16 17.75