CARPATHIAN J. MATH.
Volume **43** (2027), No. 1,
Pages 183-198

Online version at https://www.carpathian.cunbm.utcluj.ro/
Print Edition: ISSN 1584 - 2851; Online Edition: ISSN 1843 - 4401
DOI: https://doi.org/10.37193/CJM.2027.01.11

A New Two-Steps Inertial Viscosity Algorithm for Convex Bilevel Optimization Problems with Application to Machine Learning

SERDAR ROZYYEV[1] AND SUTHEP SUANTAI[2]

ABSTRACT. In this work, we propose and analyze a new accelerated fixed-point algorithm for solving convex bilevel optimization problems in real Hilbert spaces. The algorithm aims to minimize a strongly convex and differentiable objective function over the solution set of an inner-level problem, which is formulated as the minimization of the sum of smooth and nonsmooth convex functions. By incorporating both inertial and viscosity terms, the proposed method improves the convergence behavior. We rigorously establish the strong convergence of the proposed algorithm under suitable conditions.

To demonstrate practical effectiveness, we apply the algorithm to machine learning tasks, including regression and classification, using benchmark datasets. Numerical experiments show that our approach outperforms existing methods such as iBiG-SAM and aiBiG-SAM, particularly in terms of accuracy and precision, highlighting its potential for real-world applications in data-driven optimization.

1. INTRODUCTION

Fixed point theory plays an important role in nonlinear analysis and optimization, particularly in the study of iterative methods for variational inequalities, convex minimization, and machine learning problems [13, 15, 20]. A notable application is convex bilevel optimization, which involves minimizing an outer objective over the set of solutions for an inner problem, a structure commonly found in fields such as machine learning, economics, healthcare and engineering [10, 11, 12].
In this work, we consider the following convex bilevel optimization problem.

Inner-level problem:

$$\min_{x\in\mathbb{R}^n} \{f(x) + g(x)\}, \tag{1.1}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is convex and differentiable with Lipschitz continuous gradient, and $g : \mathbb{R}^n \to \mathbb{R}$ is a proper, lower semicontinuous, convex function.

Let $\Omega := \arg\min_{x\in\mathbb{R}^n}\{f(x) + g(x)\}$ denote the solution set of the inner problem.

Outer-level problem:

$$\min_{x\in\Omega} \omega(x), \tag{1.2}$$

where $\omega : \mathbb{R}^n \to \mathbb{R}$ is a strongly convex and differentiable function.

The inner-level problem can be reformulated as a monotone inclusion of the form

$$0 \in \nabla f(x) + \partial g(x),$$

Received: 13.04.2026. In revised form: 02.06.2026. Accepted: 05.08.2026
2020 *Mathematics Subject Classification.* 47H09, 47J25, 90C25, 65K05.
Key words and phrases. *Convex bilevel optimization; fixed point methods; inertial viscosity algorithm; strong convergence; machine learning.*
Corresponding author: Suthep Suantai; suthep.s@cmu.ac.th

which characterizes the optimality condition for the composite convex minimization problem [1, 13, 25]. A widely used approach for solving such problems is the forward–backward splitting method, which is discussed in Parikh and Boyd [1]. This method expresses the solution as a fixed point of the proximal-gradient operator:

$$T := \mathrm{prox}_{\alpha g}(I - \alpha \nabla f),$$

where $\alpha > 0$ is a step size. Here, $\mathrm{prox}_{\alpha g}$ denotes the proximal operator of g, defined by

$$\mathrm{prox}_{\alpha g}(u) = \arg\min_{y \in \mathbb{R}^n} \left\{ g(y) + \frac{1}{2\alpha}\|y - u\|^2 \right\}.$$

A point $\hat{x} \in \mathbb{R}^n$ solves the inner-level problem if and only if it satisfies the fixed point condition $\hat{x} = T(\hat{x})$; see, for example, [1, 10].

We denote by $\mathrm{Fix}(T)$ the set of all fixed points of T. If $\alpha \in \left(0, \frac{2}{L_f}\right)$, where L_f is the Lipschitz constant of ∇f, then T is nonexpansive and

$$\mathrm{Fix}(T) = \arg\min_{x \in \mathbb{R}^n} \{f(x) + g(x)\};$$

see, for example, [1, 10, 20].

The outer-level problem is to find a point $x^* \in \mathrm{Fix}(T)$ such that

$$\langle \nabla \omega(x^*), v - x^* \rangle \geq 0, \quad \forall v \in \mathrm{Fix}(T),$$

which characterizes the unique solution of the bilevel optimization problem [10, 13].

Let H be a real Hilbert space with inner product $\langle \cdot, \cdot \rangle$ and norm $\| \cdot \|$. A mapping $T : H \to H$ is said to be *nonexpansive* if

$$\|Tx - Ty\| \leq \|x - y\|, \quad \forall x, y \in H,$$

and a mapping $F : H \to H$ is called a *contraction* if there exists $\delta \in [0, 1)$ such that

$$\|F(x) - F(y)\| \leq \delta \|x - y\|, \quad \forall x, y \in H.$$

Over the years, many iteration methods have been developed to find a fixed point of nonexpansive mappings. A classical example is the Mann iteration[2],

$$x_{n+1} = \alpha_n x_n + (1 - \alpha_n) T x_n, \ n \in \mathbb{N}, \tag{1.3}$$

which converges weakly under suitable assumptions for nonexpansive mappings; see also [13, 15]. To obtain strong convergence, Halpern [3] introduced the iteration:

$$x_{n+1} = \alpha_n x_0 + (1 - \alpha_n) T x_n, \ n \in \mathbb{N}. \tag{1.4}$$

Later, Moudafi [4] proposed the viscosity approximation method

$$x_{n+1} = \alpha_n F(x_n) + (1 - \alpha_n) T x_n, \ n \geq 0, \tag{1.5}$$

where F is a contraction. Under appropriate control conditions, this approach ensures strong convergence to a particular fixed point selected by the contraction [4].

In this context, an inertial method uses information from previous iterates to accelerate convergence, while a viscosity approximation method employs a contraction mapping to select a particular fixed point and obtain strong convergence. Motivated by these developments, various multi-step and inertial techniques have been introduced to improve the convergence behavior of iterative methods. Ishikawa [5] and Noor [26] extended classical schemes by incorporating additional steps, while Phuengrattana and Suantai [6] proposed the SP-iteration and showed that it has a better convergence behavior than several earlier methods under suitable assumptions.

The use of inertial extrapolation, originally introduced by Polyak [8], has become a powerful tool for accelerating iterative algorithms [9, 11, 12, 23]. This idea was successfully applied by Beck and Teboulle [9] in the Fast Iterative Shrinkage-Thresholding Algorithm (FISTA), where an improved convergence rate of $O(1/k^2)$ was established for a class of convex composite optimization problems.

In the context of bilevel optimization, Sabach and Shtern [10] introduced the Bi-Level Gradient Sequential Averaging Method (BiG-SAM), which combines proximal-gradient updates for the inner problem with gradient steps for the outer problem. To enhance this framework, Shehu et al. [11] proposed the inertial variant iBiG-SAM, and later Duan and Zhang [12] developed aiBiG-SAM, incorporating alternating inertial steps to further accelerate convergence.

Although these methods have contributed significantly to the development of bilevel optimization algorithms, the choice of inertial parameters remains delicate, since large inertial parameters may accelerate the iterative process but can also affect stability. This motivates the search for alternative inertial-viscosity methods with flexible control parameters.

Motivated by these developments, this work proposes a new accelerated fixed point algorithm incorporating both inertial and viscosity mechanisms, specifically designed to solve convex bilevel optimization problems. Unlike existing approaches such as BiG-SAM and its inertial variants, the proposed method introduces a two-step inertial structure combined with a viscosity framework, which is intended to enhance the convergence behavior and provide greater flexibility in the choice of control parameters. We establish a strong convergence theorem for the generated sequence under suitable conditions, improving and generalizing several known results in the literature. Furthermore, the proposed algorithm is applied to data classification problems in machine learning, where numerical experiments demonstrate its effectiveness in terms of accuracy and efficiency.

The structure of the paper is as follows: Section 2 provides essential definitions and preliminary results used in the analysis. Section 3 presents the proposed algorithm along with a strong convergence theorem. Section 4 demonstrates the application of the algorithm to data classification tasks in machine learning. Finally, Section 5 offers concluding remarks and potential directions for future research.

2. PRELIMINARIES

In this section, we recall some useful definitions, lemmas, and known results in real Hilbert spaces that will be used throughout this work. Let $\mathbb{R}$ denote the set of real numbers, $\mathbb{R}_{\geq 0}$ the set of all nonnegative real numbers and $\mathbb{N}$ the set of natural numbers. Let H be a real Hilbert space equipped with the inner product $\langle \cdot, \cdot \rangle$ and the induced norm $\|\cdot\|$. We use $x_n \to x^*$ to denote strong convergence and $x_n \rightharpoonup x^*$ for weak convergence of a sequence $\{x_n\}$ to a point x^*, respectively.

Definition 2.1. *A mapping $T : H \to H$ is said to be*

- Lipschitzian *if there exists $\tau > 0$ such that*

$$\|Tx - Ty\| \leq \tau \|x - y\|, \quad \forall x, y \in H,$$

- a contraction *if $\tau \in [0, 1)$,*
- nonexpansive *if $\tau = 1$.*

A point $x \in H$ is a fixed point of a mapping $T : H \to H$ if $Tx = x$. The set of fixed points is denoted by $\mathrm{Fix}(T) := \{x \in H : Tx = x\}$.

Let C be a nonempty closed convex subset of a Hilbert space H. The metric projection $P_C x$ is defined by

$$P_C x = \arg\min_{y \in C} \|y - x\|.$$

Proposition 2.1 ([13]). *Let $C \subset H$ be nonempty and convex. Then,*

$$x^* = P_C x \quad \Leftrightarrow \quad \langle x - x^*, y - x^* \rangle \leq 0, \quad \forall y \in C.$$

Definition 2.2. *A function $f : \mathbb{R}^n \to \mathbb{R}$ is strongly convex with constant $\sigma > 0$ if for any $x, y \in \mathbb{R}^n$ and $\lambda \in [0, 1]$,*

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y) - \frac{\sigma}{2}\lambda(1-\lambda)\|x - y\|^2.$$

If f is differentiable, this is equivalent to:

$$f(y) - f(x) \geq \langle \nabla f(x), y - x \rangle + \frac{\sigma}{2}\|y - x\|^2.$$

Let $\{T_n\}$ and Λ be families of nonexpansive mappings from a real Hilbert space H into itself, such that $\varnothing \neq F(\Lambda) \subset \Gamma := \bigcap_{n=1}^{\infty} Fix(T_n)$, where $F(\Lambda)$ is the set of common fixed points of all mappings in Λ. We say that $\{T_n\}$ satisfy the *NST-condition(I)* with respect to Λ [18, 7], if for every bounded sequence $\{x_n\} \subset H$,

$$\lim_{n\to\infty} \|x_n - T_n x_n\| = 0 \;\Rightarrow\; \lim_{n\to\infty} \|x_n - T x_n\| = 0, \;\; \forall T \in \Gamma.$$

If $\Gamma = \{T\}$ is a singleton, then $\{T_n\}$ is said to satisfy the *NST-condition(I)* with respect to T.

Definition 2.3 ([19, 14]). *Let $\{T_n : H \to H\}$ be a sequence of nonexpansive mappings with a nonempty common fixed point set*

$$\Gamma := \bigcap_{n=1}^{\infty} \mathrm{Fix}(T_n).$$

*The sequence $\{T_n\}$ satisfies **condition (Z)** if for every bounded sequence $\{x_n\} \subset H$ such that $\lim_{n\to\infty} \|x_n - T_n x_n\| \to 0$, every weak cluster point of $\{x_n\}$ belongs to Γ. These two concepts were introduced by Takahashi*[19] *and Aoyama* [14]

Definition 2.4 ([15]). *Let $T : H \to H$ be a mapping. The operator $I - T$ is said to be demiclosed at zero if for any sequence $\{x_n\} \subset H$ with $x_n \rightharpoonup x$ and $\|x_n - T x_n\| \to 0$, then $x \in \mathrm{Fix}(T)$.*

It is well known that if T is nonexpansive, then $I - T$ is demiclosed at zero.

Remark 2.1. *If $\{T_n\}$ satisfies NST - condition (I) with respect to T, then $\{T_n\}$ satisfies condition (Z). To see this, suppose $\{x_n\}$ is a bounded sequence such that*

$$\lim_{n\to\infty} \|x_n - T_n x_n\| = 0.$$

By the NST - condition (I) with respect to T, we have

$$\lim_{n\to\infty} \|x_n - T x_n\| = 0.$$

Now, let $x \in \omega(\{x_n\})$. Then there is a subsequence $\{x_{n_k}\}$ of $\{x_n\}$ such that $x_{n_k} \rightharpoonup x$. It follows by the demiclosedness of $I - T$, we obtain $x \in \mathrm{Fix}(T) \subset \Gamma = \bigcap_{n=1}^{\infty} \mathrm{Fix}(T_n)$. That is, $\{T_n\}$ satisfies the condition (Z).

Lemma 2.1 (Opial's Lemma). *Let H be a real Hilbert space and let $\{x_n\} \subset H$. Suppose that*

(i) for every $x \in H$, the limit $\lim_{n\to\infty} \|x_n - x\|$ exists;

(ii) every weak sequential cluster point of $\{x_n\}$ belongs to a nonempty subset $\Gamma \subset H$.

Then there exists a unique point $x^* \in \Gamma$ *such that*

$$x_n \rightharpoonup x^*.$$

In particular, if $\Gamma = \mathrm{Fix}(T)$ *for some mapping* $T : H \to H$*, then*

$$x_n \rightharpoonup x^* \in \mathrm{Fix}(T).$$

Theorem 2.1 (**Banach Fixed Point Theorem**). *If* $T : C \to C$ *is a contraction on a nonempty closed subset* $C \subset H$*, then* T *has a unique fixed point in* C*.*

Definition 2.5. *Let* H *be a real Hilbert space. Let* $g : H \to (-\infty, +\infty]$ *be a proper, convex, and lower semi-continuous function, and let* $f : H \to \mathbb{R}$ *be a convex and differentiable function with L-Lipschitz continuous gradient for some* $L > 0$*. For a stepsize* $\lambda \in (0, 2/L)$*, the forward–backward operator associated with* f *and* g *is defined by*

$$T = \mathrm{prox}_{\lambda g}(x - \lambda \nabla f(x)), \qquad x \in H,$$

where the proximal mapping of g *is given by*

$$\mathrm{prox}_{\lambda g}(x) = \arg\min_{y \in H} \left\{ g(y) + \frac{1}{2\lambda} \|y - x\|^2 \right\}.$$

Lemma 2.2 ([20]). *Let* $g : H \to \mathbb{R} \cup \{+\infty\}$ *be a proper, convex, and lower semi-continuous function, and let* $f : H \to \mathbb{R}$ *be a convex and differentiable function whose gradient* ∇f *is L-Lipschitz continuous for some* $L > 0$*. If* $\{T_n\}$ *is the forward–backward operator of* f *and* g*, with respect to step sizes* $c_n \in (0, 2/L)$*, such that* $\{c_n\}$ *converges to* c*, then* $\{T_n\}$ *satisfies the NST-condition (I) with T, where T is the forward–backward operator of f and g with respect to* $c \in (0, 2/L)$*.*

Proposition 2.2 ([10]). *Let* $\omega : \mathbb{R}^n \to \mathbb{R}$ *be strongly convex with constant* $\sigma > 0$*, and assume* $\nabla\omega$ *is* L_ω*-Lipschitz. Then, the mapping* $S_s = I - s\nabla\omega$ *is a contraction for* $0 < s \leq 2/(L_\omega + \sigma)$*, that is*

$$\|x - s\nabla\omega(x) - (y - s\nabla\omega(y))\| \leq \sqrt{1 - \frac{2s\sigma L_\omega}{\sigma + L_\omega}}\, \|x - y\|, \quad \forall x, y \in \mathbb{R}^n.$$

Lemma 2.3 ([13],[21]). *For* $x, y \in H$ *and* $t \in [0, 1]$*, the following hold:*

(i) $\|tx + (1-t)y\|^2 = t\|x\|^2 + (1-t)\|y\|^2 - t(1-t)\|x - y\|^2$,
(ii) $\|x \pm y\|^2 = \|x\|^2 \pm 2\langle x, y\rangle + \|y\|^2$,
(iii) $\|x + y\|^2 \leq \|x\|^2 + 2\langle y, x + y\rangle$.

Lemma 2.4 ([22]). *Assume that* $\{a_n\} \subset \mathbb{R}_{\geq 0}$*,* $\{b_n\} \subset \mathbb{R}$*, and* $\{t_n\} \subset (0, 1)$ *with* $\sum_{n=1}^{\infty} t_n = \infty$*. Assume that*

$$a_{n+1} \leq (1 - t_n)a_n + t_n b_n.$$

If $\limsup_{i\to\infty} b_{n_i} \leq 0$ *for every subsequence* $\{a_{n_i}\}$ *of* $\{a_n\}$ *satisfying*

$$\liminf_{i\to\infty} (a_{n_i+1} - a_{n_i}) \geq 0.$$

Then $a_n \to 0$ *as* $n \to \infty$*.*

3. Main Results

In this section, based on the viscosity approximation method, we introduce a new accelerated algorithm using the inertial technique to find a common fixed point of a countable family of nonexpansive operators satisfying the condition (Z) in a Hilbert space. A strong convergence theorem is proved under some suitable control conditions. After that, we apply the proposed algorithm to solve a convex minimization problem of the sum of two convex functions.

We begin by setting the following hypotheses.

- $\mathcal{H}$ is a real Hilbert space;
- $\{T_n : \mathcal{H} \to \mathcal{H}\}$ is a family of nonexpansive operators;
- $\Gamma := \bigcap_{n=1}^{\infty} \mathrm{Fix}(T_n) \neq \emptyset$;
- $\{T_n\}$ satisfies the condition (Z);
- $F : \mathcal{H} \to \mathcal{H}$ is a k-contraction, where $0 < k < 1$.

In optimization theory, inertial-type extrapolation is commonly used to accelerate iterative methods by adding a momentum term such as $\theta_n(x_n - x_{n-1})$ [8, 23]. The proposed method is designed by combining three components: an inertial extrapolation step, a viscosity approximation step, and a second inertial correction step. The term $\theta_n(x_n - x_{n-1})$ uses information from previous iterates to accelerate the process, while the viscosity term involving the contraction F helps select a particular fixed point and obtain strong convergence. The final inertial term $\alpha_n(y_n - y_{n-1})$ provides an additional correction and increases the flexibility of the iterative scheme. Compared with iBiG-SAM and aiBiG-SAM, the proposed method uses a two-step inertial-viscosity structure within a fixed-point framework. Now, we are ready to introduce a modified inertial viscosity approximation method as follows:

Algorithm 1. A Modified Inertial Viscosity Iterative Algorithm (MIVIA)

Initialization: Let $\{\beta_n\}$, $\{\gamma_n\}$, $\{\xi_n\}$, $\{\tau_n\}$ be sequences of positive real numbers and let $\{\eta_n\}$ and $\{\mu_n\}$ be bounded sequences of nonnegative real numbers. Take $x_0, x_1, y_0 \in \mathcal{H}$ arbitrarily.

Iterative steps: For $n \geq 1$, calculate x_{n+1} as follows:

Step 1. Compute an inertial parameter

$$\theta_n = \begin{cases} \min\left\{\mu_n, \dfrac{\tau_n}{\|x_n - x_{n-1}\|}\right\}, & \text{if } x_n \neq x_{n-1}, \\ \mu_n, & \text{otherwise.} \end{cases}$$

Step 2. Compute

$$\begin{cases} w_n &= x_n + \theta_n(x_n - x_{n-1}), \\ z_n &= (1-\gamma_n)T_n w_n + \gamma_n F(w_n), \\ y_n &= (1-\beta_n)T_n w_n + \beta_n T_n z_n, \\ x_{n+1} &= y_n + \alpha_n(y_n - y_{n-1}). \end{cases}$$

where

$$\alpha_n = \begin{cases} \min\left\{\eta_n, \dfrac{\xi_n}{\|y_n - y_{n-1}\|}\right\}, & \text{if } y_n \neq y_{n-1}, \\ \eta_n, & \text{otherwise.} \end{cases}$$

Set $n := n + 1$ and return to Step 1.

For practical implementation in data classification, the performance of the above algorithm was evaluated using accuracy, precision, and recall as the primary metrics. The experiment was conducted iteratively, with the termination criterion defined by a maximum number of iterations. At each iteration, the model's classification performance was assessed based on the aforementioned metrics to determine convergence, and the experiment was terminated once either the performance metrics stabilized at their optimal level or the maximum iteration limit was reached.

A strong convergence result is established by the following theorem.

Theorem 3.2. *Any sequence $\{x_n\}$ generated by Algorithm 1 converges strongly to an element $x^* \in \Gamma$, where $x^* = P_\Gamma F(x^*)$, provided that the sequences $\{\beta_n\}$, $\{\gamma_n\}$, $\{\xi_n\}$ and $\{\tau_n\}$ satisfy the following conditions:*

(C1) $0 < \epsilon_1 \leq \beta_n \leq \epsilon_2 < 1$ *for some* $\epsilon_1, \epsilon_2 \in \mathbb{R}$
(C2) $0 < \gamma_n < 1$, $\lim_{n\to\infty} \gamma_n = 0$ *and* $\sum_{n=1}^{\infty} \gamma_n = \infty$;
(C3) $\lim_{n\to\infty} \frac{\tau_n}{\gamma_n} = 0$;
(C4) $\lim_{n\to\infty} \frac{\xi_n}{\gamma_n} = 0$.

Proof. Since the operator $P_\Gamma F$ is a contraction, the Banach fixed point theorem ensures the existence and uniqueness of a point $x^* \in \Gamma$ satisfying $x^* = P_\Gamma F(x^*)$. It follows that $x^* \in \mathrm{Fix}(T_n)$ for all $n \in \mathbb{N}$.

We begin by showing that the sequence $\{x_n\}$ is bounded. From the definition of w_n we have:

$$\|w_n - x^*\| \leq \|x_n - x^*\| + \theta_n \|x_n - x_{n-1}\|.$$

Next, considering z_n, and using the nonexpansiveness of T_n with the contraction property of F

$$\begin{aligned}
\|z_n - x^*\| &\leq (1-\gamma_n)\|T_n w_n - x^*\| + \gamma_n \|F(w_n) - x^*\| \\
&\leq (1-\gamma_n)\|T_n w_n - x^*\| + \gamma_n \|F(w_n) - F(x^*)\| + \gamma_n \|F(x^*) - x^*\| \\
&\leq (1-\gamma_n)\|w_n - x^*\| + \gamma_n k \|w_n - x^*\| + \gamma_n \|F(x^*) - x^*\| \\
&= (1-\gamma_n(1-k))\|w_n - x^*\| + \gamma_n \|F(x^*) - x^*\|,
\end{aligned}$$

Similarly,

$$\begin{aligned}
\|y_n - x^*\| &\leq (1-\beta_n)\|T_n w_n - x^*\| + \beta_n \|T_n z_n - x^*\| \\
&\leq (1-\beta_n)\|w_n - x^*\| + \beta_n \|z_n - x^*\| \\
&\leq (1-\beta_n)\|w_n - x^*\| + \beta_n (1-\gamma_n(1-k))\|w_n - x^*\| + \beta_n \gamma_n \|F(x^*) - x^*\| \\
&= (1-\beta_n\gamma_n(1-k))\|w_n - x^*\| + \beta_n \gamma_n \|F(x^*) - x^*\| \\
&\leq (1-\beta_n\gamma_n(1-k))(\|x_n - x^*\| + \theta_n \|x_n - x_{n-1}\|) + \beta_n \gamma_n \|F(x^*) - x^*\|,
\end{aligned}$$

Then we obtain

$$\begin{aligned}
\|x_{n+1} - x^*\| &\leq \|y_n - x^*\| + \alpha_n \|y_n - y_{n-1}\| \\
&\leq (1-\beta_n\gamma_n(1-k))(\|x_n - x^*\| + \theta_n \|x_n - x_{n-1}\|) + \beta_n \gamma_n \|F(x^*) - x^*\| \\
&\quad + \alpha_n \|y_n - y_{n-1}\| \\
&\leq (1-\beta_n\gamma_n(1-k))\|x_n - x^*\| \\
&\quad + \beta_n\gamma_n \left(\frac{\theta_n}{\beta_n\gamma_n}\|x_n - x_{n-1}\| + \|F(x^*) - x^*\| + \frac{\alpha_n}{\beta_n\gamma_n}\|y_n - y_{n-1}\| \right)
\end{aligned}$$

By the conditions of the sequences $\{\theta_n\}$ and $\{\alpha_n\}$, and from (C3) and (C4), we have

$$\lim_{n\to\infty} \frac{\theta_n}{\beta_n\gamma_n}\|x_n - x_{n-1}\| = 0 \ \text{ and } \ \lim_{n\to\infty} \frac{\alpha_n}{\beta_n\gamma_n}\|y_n - y_{n-1}\| = 0.$$

Therefore, there exist a constant $M > 0$ such that

$$\frac{\theta_n}{\beta_n\gamma_n}\|x_n - x_{n-1}\| \leq M \ \text{ and } \ \frac{\alpha_n}{\beta_n\gamma_n}\|y_n - y_{n-1}\|] \leq M \ \text{ for all } \ n \in \mathbb{N}.$$

Hence,

$$\begin{aligned}\|x_{n+1}-x^*\| &\leq (1-\beta_n\gamma_n(1-k))\|x_n-x^*\|+\beta_n\gamma_n(1-k)(\frac{2M+\|F(x^*)-x^*\|}{1-k})\\ &\leq \max\left\{\|x_n-x^*\|, \frac{2M+\|F(x^*)-x^*\|}{1-k}\right\}.\end{aligned}$$

By mathematical induction, we deduce that

$$\|x_n-x^*\| \leq \max\left\{\|x_1-x^*\|, \frac{2M+\|F(x^*)-x^*\|}{1-k}\right\}, \forall n \in \mathbb{N}.$$

This implies that the sequence $\{x_n\}$ is bounded. Consequently, the sequences $\{w_n\}$,$\{z_n\}$,$\{y_n\}$, $\{T_nw_n\}$,$\{T_nz_n\}$ and $\{F(w_n)\}$ are also bounded.
By Lemma 2.3 (ii), we have

$$\begin{aligned}\|w_n-x^*\|^2 &= \|x_n-x^*+\theta_n(x_n-x_{n-1})\|^2\\ &= \|x_n-x^*\|^2+\theta_n^2\|x_n-x_{n-1}\|^2+2\theta_n\langle x_n-x^*, x_n-x_{n-1}\rangle\\ &\leq \|x_n-x^*\|^2+\theta_n^2\|x_n-x_{n-1}\|^2+2\theta_n\|x_n-x^*\|\|x_n-x_{n-1}\|. \qquad (3.6)\end{aligned}$$

By Lemma 2.3 (i) and (iii) we have

$$\begin{aligned}\|z_n-x^*\|^2 &= \|(1-\gamma_n)(T_nw_n-x^*)+\gamma_n(F(w_n)-F(x^*))+\gamma_n(F(x^*)-x^*))\|^2\\ &\leq \|(1-\gamma_n)(T_nw_n-x^*)+\gamma_n(F(w_n)-F(x^*))\|^2+2\gamma_n\langle F(x^*)-x^*, z_n-x^*\rangle\\ &\leq (1-\gamma_n)\|T_nw_n-x^*\|^2+\gamma_n\|F(w_n)-F(x^*)\|^2+2\gamma_n\langle F(x^*)-x^*, z_n-x^*\rangle\\ &\leq (1-\gamma_n)\|w_n-x^*\|^2+k\gamma_n\|w_n-x^*\|^2+2\gamma_n\langle F(x^*)-x^*, z_n-x^*\rangle\\ &\leq (1-\gamma_n(1-k))\|w_n-x^*\|^2+2\gamma_n\langle F(x^*)-x^*, z_n-x^*\rangle \qquad (3.7)\end{aligned}$$

By Lemma 2.3 (i) together with (3.6) and (3.7), we get

$$\begin{aligned}\|y_n-x^*\|^2 &= \|(1-\beta_n)(T_nw_n-x^*)+\beta_n(T_nz_n-x^*)\|^2\\ &= (1-\beta_n)\|T_nw_n-x^*\|^2+\beta_n\|T_nz_n-x^*\|^2-\beta_n(1-\beta_n)\|T_nw_n-T_nz_n\|^2\\ &\leq (1-\beta_n)\|w_n-x^*\|^2+\beta_n\|z_n-x^*\|^2-\beta_n(1-\beta_n)\|T_nw_n-T_nz_n\|^2\\ &\leq (1-\beta_n)\|w_n-x^*\|^2+\beta_n((1-\gamma_n(1-k))\|w_n-x^*\|^2+2\gamma_n\langle F(x^*)-x^*, z_n-x^*\rangle)\\ &\quad -\beta_n(1-\beta_n)\|T_nw_n-T_nz_n\|^2\\ &\leq (1-\beta_n\gamma_n(1-k))\|w_n-x^*\|^2+2\beta_n\gamma_n\langle F(x^*)-x^*, z_n-x^*\rangle\\ &\quad -\beta_n(1-\beta_n)\|T_nw_n-T_nz_n\|^2 \qquad (3.8)\end{aligned}$$

Using the definitions of x_{n+1}, together with (3.8) and Lemma 2.3 (ii), we obtain

$$\begin{aligned}
\|x_{n+1}-x^*\|^2 &= \|y_n - x^* + \alpha_n(y_n - y_{n-1})\|^2 \\
&= \|y_n - x^*\|^2 + \alpha_n^2\|y_n - y_{n-1})\|^2 + 2\alpha_n\langle y_n - x^*, y_n - y_{n-1}\rangle \\
&\leq \|y_n - x^*\|^2 + \alpha_n^2\|y_n - y_{n-1})\|^2 + 2\alpha_n\|y_n - x^*\|\|y_n - y_{n-1}\| \\
&\leq (1-\beta_n\gamma_n(1-k))\|w_n - x^*\|^2 + 2\beta_n\gamma_n\langle F(x^*) - x^*, z_n - x^*\rangle \\
&\quad - \beta_n(1-\beta_n)\|T_nw_n - T_nz_n\|^2 + \alpha_n^2\|y_n - y_{n-1}\|^2 + 2\alpha_n\|y_n - x^*\|\|y_n - y_{n-1}\| \\
&\leq (1-\beta_n\gamma_n(1-k))(\|x_n - x^*\|^2 + \theta_n^2\|x_n - x_{n-1}\|^2 + 2\theta_n\|x_n - x^*\|\|x_n - x_{n-1}\|) \\
&\quad + 2\beta_n\gamma_n\langle F(x^*) - x^*, z_n - x^*\rangle - \beta_n(1-\beta_n)\|T_nw_n - T_nz_n\|^2 \\
&\quad + \alpha_n^2\|y_n - y_{n-1}\|^2 + 2\alpha_n\|y_n - x^*\|\|y_n - y_{n-1}\| \\
&\leq (1-\beta_n\gamma_n(1-k))\|x_n - x^*\|^2 + \theta_n^2\|x_n - x_{n-1}\|^2 + 2\theta_n\|x_n - x^*\|\|x_n - x_{n-1}\| \\
&\quad + \alpha_n^2\|y_n - y_{n-1})\|^2 + 2\alpha_n\|y_n - x^*\|\|y_n - y_{n-1}\| \\
&\quad + 2\beta_n\gamma_n\langle F(x^*) - x^*, z_n - x^*\rangle - \beta_n(1-\beta_n)\|T_nw_n - T_nz_n\|^2 \\
&= (1-\beta_n\gamma_n(1-k))\|x_n - x^*\|^2 - \beta_n(1-\beta_n)\|T_nw_n - T_nz_n\|^2 + \beta_n\gamma_n(1-k)b_n, \qquad (3.9)
\end{aligned}$$

where

$$\begin{aligned}
b_n = \frac{1}{1-k}\Bigg\{ & 2\langle F(x^*) - x^*, z_n - x^*\rangle + \theta_n\|x_n - x_{n-1}\|\left(\frac{\theta_n}{\beta_n\gamma_n}\|x_n - x_{n-1}\|\right) \\
& + 2\|x_n - x^*\|\left(\frac{\theta_n}{\beta_n\gamma_n}\|x_n - x_{n-1}\|\right) + \alpha_n\|y_n - y_{n-1}\|\left(\frac{\alpha_n}{\beta_n\gamma_n}\|y_n - y_{n-1}\|\right) \\
& + 2\|y_n - x^*\|\left(\frac{\alpha_n}{\beta_n\gamma_n}\|y_n - y_{n-1}\|\right)\Bigg\}.
\end{aligned}$$

From (3.9), we get

$$\beta_n(1-\beta_n)\|T_nw_n - T_nz_n\|^2 \leq \|x_n - x^*\|^2 - \|x_{n+1} - x^*\|^2 + \beta_n\gamma_n(1-k)M_1, \qquad (3.10)$$

where $M_1 = \sup\{b_n : n \in \mathbb{N}\}$.

Now, set $a_n := \|x_n - x^*\|^2$, $\quad t_n := \beta_n\gamma_n(1-k)$, from (3.9), we get

$$a_{n+1} \leq (1-t_n)a_n + t_nb_n.$$

We aim to show that $a_n \to 0$ by using Lemma 2.4. To show this, suppose that $\{a_{n_i}\}$ is a subsequence of $\{a_n\}$ such that

$$\liminf_{i\to\infty}(a_{n_i+1} - a_{n_i}) \geq 0.$$

Then, by inequality (3.10), we have

$$\beta_n(1-\beta_n)\|T_nw_n - T_nz_n\|^2 \leq a_n - a_{n+1} + \beta_n\gamma_n(1-k)M_1,$$

Now, taking $\limsup$ over a subsequence $\{n_i\}$ and applying the condition (C2), we obtain:

$$\begin{aligned}
\limsup_{i\to\infty}\beta_{n_i}(1-\beta_{n_i})\|T_{n_i}w_{n_i} - T_{n_i}z_{n_i}\|^2 &\leq \limsup_{i\to\infty}(a_{n_i} - a_{n_i+1} + \beta_{n_i}\gamma_{n_i}(1-k)M_1) \\
&\leq \limsup_{i\to\infty}(a_{n_i} - a_{n_i+1}) + (1-k)M_1\lim_{i\to\infty}\beta_{n_i}\gamma_{n_i} \\
&= -\liminf_{i\to\infty}(a_{n_i+1} - a_{n_i}) \\
&\leq 0.
\end{aligned}$$

This implies by (C1) that

$$\lim_{i\to\infty}\|T_{n_i}w_{n_i} - T_{n_i}z_{n_i}\| = 0. \qquad (3.11)$$

From the definition of the algorithm, we get

$$\|z_{n_i} - T_{n_i} w_{n_i}\| = \gamma_{n_i} \|F(w_{n_i}) - T_{n_i} w_{n_i}\|.$$

This implies by (C2) that

$$\lim_{i\to\infty} \|z_{n_i} - T_{n_i} w_{n_i}\| = 0. \tag{3.12}$$

From

$$\|z_{n_i} - T_{n_i} z_{n_i}\| \leq \|z_{n_i} - T_{n_i} w_{n_i}\| + \|T_{n_i} w_{n_i} - T_{n_i} z_{n_i}\|,$$

it follows by (3.11) and (3.12) that

$$\lim_{i\to\infty} \|z_{n_i} - T_{n_i} z_{n_i}\| = 0.$$

We now show that

$$\limsup_{i\to\infty} \langle F(x^*) - x^*, z_{n_i} - x^* \rangle \leq 0.$$

Since $\{z_{n_i}\}$ is bounded, we can choose a subsequence $\{z_{n_{i_j}}\}$ of $\{z_{n_i}\}$ such that $z_{n_{i_j}} \rightharpoonup z \in H$ and

$$\lim_{i\to\infty} \langle F(x^*) - x^*, z_{n_{i_j}} - x^* \rangle = \limsup_{i\to\infty} \langle F(x^*) - x^*, z_{n_i} - x^* \rangle.$$

Since $\{T_n\}$ satisfies the condition (Z) and $\lim_{i\to\infty} \|z_{n_i} - T_{n_i} z_{n_i}\| = 0$, we get $z \in \Gamma$. Moreover, since $z \in \Gamma$ and $x^* = P_\Gamma(F(x^*))$, by Proposition 2.1, we have

$$\langle F(x^*) - x^*, z - x^* \rangle \leq 0.$$

It follows that

$$\limsup_{i\to\infty} \langle F(x^*) - x^*, z_{n_i} - x^* \rangle = \lim_{j\to\infty} \langle F(x^*) - x^*, z_{n_{i_j}} - x^* \rangle = \lim_{i\to\infty} \langle F(x^*) - x^*, z - x^* \rangle \leq 0.$$

Consequently, we deduce that

$$\limsup_{i\to\infty} b_{n_i} \leq 0.$$

By Lemma 2.4, we can conclude that $\alpha_n \to 0$ as $n \to \infty$. That is, $x_n \to x^*$ as $n \to \infty$. The proof is complete. □

Next, we present an algorithm (Algorithm 2) for solving a convex bilevel optimization problem. Algorithm 2 is obtained from Algorithm 1 by choosing T_n as the forward–backward operator

$$T_n = \operatorname{prox}_{c_n g}(I - c_n \nabla f)$$

associated with the inner-level problem, and by choosing the contraction mapping $F = I - s\nabla\omega$ associated with the outer-level objective. Thus, Algorithm 2 is a concrete forward–backward implementation of Algorithm 1 for the convex bilevel optimization problem.

Algorithm 2. Modified Inertial Forward-Backward Algorithm (MIFBA)

Initialization: Let $\{\beta_n\}$, $\{\gamma_n\}$, $\{\xi_n\}$ and $\{\tau_n\}$ be sequences of positive real numbers and let $\{\eta_n\}$ and $\{\mu_n\}$ be bounded sequences of nonnegative real numbers.
Let

$$\{c_n\} \subset \left(0, \frac{2}{L_f}\right) \quad \text{with} \quad \lim_{n\to\infty} c_n = c \in \left(0, \frac{2}{L_\omega}\right) \text{and } 0 < s \leq \frac{2}{L_\omega + \sigma}$$

Take $x_0, x_1, y_0 \in \mathcal{H}$ arbitrarily.
Iterative steps: For $n \geq 1$, calculate x_{n+1} as follows:
Step 1. Compute an inertial parameter

$$\theta_n = \begin{cases} \min\left\{\mu_n, \dfrac{\tau_n}{\|x_n - x_{n-1}\|}\right\}, & \text{if } x_n \neq x_{n-1}, \\ \mu_n, & \text{otherwise.} \end{cases}$$

Step 2. Compute

$$\begin{cases} w_n & = x_n + \theta_n(x_n - x_{n-1}), \\ z_n & = (1-\gamma_n)\text{prox}_{c_n g}\left(x - c_n\nabla f\right) w_n + \gamma_n(I - s\nabla\omega), \\ y_n & = (1-\beta_n)\text{prox}_{c_n g}\left(x - c_n\nabla f\right) w_n + \beta_n \text{prox}_{c_n g}\left(x - c_n\nabla f\right) z_n, \\ x_{n+1} & = y_n + \alpha_n(y_n - y_{n-1}). \end{cases}$$

where

$$\alpha_n = \begin{cases} \min\left\{\eta_n, \dfrac{\xi_n}{\|y_n - y_{n-1}\|}\right\}, & \text{if } y_n \neq y_{n-1}, \\ \eta_n, & \text{otherwise.} \end{cases}$$

Set $n := n + 1$ and return to Step 1.

For practical implementation, the iterative process is terminated when a prescribed maximum number of iterations $N_{\max}$ is reached.

Theorem 3.3. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex and continuously differentiable function whose gradient ∇f is Lipschitz continuous with constant $L_f > 0$. Let $g : \mathbb{R}^n \to \mathbb{R}$ be a proper, convex, and lower semicontinuous function.*

Assume that the solution set of the inner problem

$$\Omega := \text{argmin}_{x\in\mathbb{R}^n}\{f(x) + g(x)\}$$

is nonempty.

Let $\omega : \mathbb{R}^n \to \mathbb{R}$ be a strongly convex and continuously differentiable function with strong convexity parameter $\sigma > 0$ and Lipschitz continuous gradient with constant $L_\omega > 0$.

Let $\{x_n\}$ be the sequence generated by Algorithm 2, under the control conditions:

(C1) $0 < \epsilon_1 \leq \beta_n \leq \epsilon_2 < 1$;
(C2) $0 < \gamma_n < 1$, $\lim_{n\to\infty}\gamma_n = 0$ *and* $\sum_{n=1}^{\infty}\gamma_n = \infty$;
(C3) $\lim_{n\to\infty}\frac{\tau_n}{\gamma_n} = 0$;
(C4) $\lim_{n\to\infty}\frac{\xi_n}{\gamma_n} = 0$.

Then the sequence $\{x_n\}$ converges strongly to the unique point $x^* \in \Omega$ satisfying

$$x^* = P_\Omega F(x^*),$$

where P_Ω denotes the metric projection onto Ω and $F = I - s\nabla\omega$.

Moreover, x^* solves the convex bi-level optimization problem

$$\min_{x\in\Omega}\omega(x).$$

Proof. Set the *proximal gradient operator*

$$T_n := \operatorname{prox}_{c_n g}(x - c_n\nabla f)$$

and the *contraction mapping*

$$F := I - s\nabla\omega.$$

Since f is convex and continuously differentiable with Lipschitz continuous gradient, and g is proper, convex, and lower semicontinuous, the proximal gradient mapping T_n is nonexpansive for each n. Moreover, by Lemma 2.2, the sequence $\{T_n\}$ satisfies the NST-condition (I) with

$$T = \operatorname{prox}_{cg}(x - c\nabla f), \quad \text{where} \quad c = \lim_{n\to\infty} c_n.$$

By Remark 2.1, the sequence $\{T_n\}$ satisfies the condition (Z).
From Proposition 2.2, we know that the mapping

$$F = I - s\nabla\omega(x)$$

is a contraction.

Thus, Algorithm 2 is precisely a particular case of Algorithm 1 applied to the mappings T_n and F.

Applying Theorem 3.2, we deduce that the sequence $\{x_n\}$ converges strongly to a point $x^* \in \Omega$ such that

$$x^* = P_\Omega F(x^*),$$

where P_Ω is the metric projection onto Ω.

Finally, by Proposition 2.1, we have:

$$x^* = P_\Omega F(x^*) \quad \Longleftrightarrow \quad \langle F(x^*) - x^*, x - x^*\rangle \le 0, \quad \forall x \in \Omega.$$

Substituting $F(x^*) = x^* - s\nabla\omega(x^*)$ gives

$$\langle -s\nabla\omega(x^*), x - x^*\rangle \le 0, \quad \forall x \in \Omega,$$

which is equivalent to

$$\langle \nabla\omega(x^*), x - x^*\rangle \ge 0, \quad \forall x \in \Omega.$$

This is precisely the first-order optimality condition for the minimization of ω over Ω.

Hence, x^* is the unique solution to the convex bilevel optimization problem

$$\min_{x\in\Omega}\omega(x),$$

that completes the proof.

□

4. Numerical Results

This section presents numerical experiments conducted to evaluate the performance of the proposed bilevel optimization algorithm, **MIVIA**, in comparison with two well-known bilevel methods: **iBiG-SAM** and **aiBiG-SAM**. The primary goal is to evaluate the effectiveness of MIVIA in solving real-world classification problems and to demonstrate its capability to achieve superior predictive performance under a unified experimental framework.

We adopt the Extreme Learning Machine (ELM) framework [24], a single-layer feedforward neural network (SLFN) known for its fast learning speed and strong generalization

performance. In ELM, the input weights and biases of the hidden layer are randomly assigned and fixed, while only the output weights are learned [24]. This design transforms the learning task into a convex optimization problem [24].

Let the training dataset be represented as

$$\{(x_k, t_k) \in \mathbb{R}^n \times \mathbb{R}^m : k = 1, 2, \ldots, s\},$$

where x_k is an input vector and t_k is its corresponding target label. Following the standard ELM formulation [24], the SLFN computes the output for sample x_k as

$$o_k = \sum_{j=1}^{h} \alpha_j \, G(\langle \omega_j, x_k \rangle + b_j), \quad k = 1, 2, \ldots, s,$$

where h denotes the number of hidden neurons, $G(\cdot)$ is the activation function, and ω_j, b_j, and α_j are the input weights, biases, and output weights, respectively.

The aim of SLFNs is to find the parameters w_i, α_i, and b_i for all $i = 1, 2, 3, \ldots, s$ such that

$$t_k = \sum_{j=1}^{h} \alpha_j \, G(\langle \omega_j, x_k \rangle + b_j), \quad k = 1, 2, \ldots, s,$$

which means the zero error $\sum_{j=1}^{h} \|o_j - t_j\|$ is close to o while ELM goal is to find only α_j with w_j and b_j chosen at random.

Let $H \in \mathbb{R}^{s \times h}$ denote the hidden-layer output matrix defined, as in the ELM framework [24], by

$$\mathbf{H} = \begin{bmatrix} G(\langle w_1, x_1 \rangle + b_1) & \cdots & G(\langle w_h, x_1 \rangle + b_h) \\ \vdots & \ddots & \vdots \\ G(\langle w_1, x_s \rangle + b_1) & \cdots & G(\langle w_h, x_s \rangle + b_h) \end{bmatrix}.$$

Then, the overall output prediction equation becomes:

$$Hu = T, \tag{4.13}$$

where $u = [\alpha_1, \alpha_2, \ldots, \alpha_h]^T$ represents the output weights and $T = [t_1, t_2, \ldots, t_s]^T$ is the matrix of target labels.

This leads to the least-squares problem

$$\min_u \|Hu - T\|^2.$$

To prevent overfitting, we incorporate ℓ_1-regularization following the LASSO framework [27], yielding

$$\min_u \|Hu - T\|^2 + \lambda \|u\|_1,$$

where $\lambda > 0$ is a regularization parameter.

This formulation naturally fits into a bilevel optimization framework:

- **Inner problem:** $\varphi(u) = \|Hu - T\|^2, \quad \psi(u) = \lambda \|u\|_1,$
- **Outer problem:** $\omega(u) = \frac{1}{2}\|u\|^2.$

Note that ω is strongly convex with parameter $\sigma_\omega = 1$ and has Lipschitz continuous gradient with constant $L_\omega = 1$.

To evaluate the proposed method, we conduct experiments on two real-world medical datasets:

- **Heart Disease Dataset:** 303 samples, 13 features from the UCI Machine Learning Repository [28].

- **Hypertension Dataset:** 6108 samples, 9 features (Sriphat Medical Center, Chiang Mai University).

In the numerical experiments, we set $N_{\max} = 1000$. Each dataset was split into 70% training and 30% testing data, and a one-hidden-layer ELM model was used.

We evaluate the classification performance using three standard metrics:

- **Accuracy:**

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Cases}} \times 100.$$

- **Precision:**

$$\text{Precision} = \frac{TP}{TP + FP},$$

where TP is the number of true positives and FP is the number of false positives. Precision is particularly important in medical applications where minimizing false alarms is critical.

- **Recall:**

$$\text{Recall} = \frac{TP}{TP + FN},$$

where TP is the number of true positives, and FN is the number of false negatives.

Recall is particularly important in contexts such as medical diagnosis, where missing a true positive case can be more costly than a false alarm. High recall ensures that the model identifies as many positive cases as possible, complementing precision by minimizing false negatives.

These metrics provide a balanced view of model performance in terms of both general correctness and reliability of positive classifications.

TABLE 1. Accuracy, Precision, and Recall Comparison on Heart Disease and Hypertension Datasets

Dataset	**Algorithm**	Accuracy (Train)	Accuracy (Test)	Precision (Train)	Precision (Test)	Recall (Train)	Recall (Test)
Heart Disease	MIVIA	85.37	90.10	0.8281	0.8596	0.9217	0.9800
	iBiG-SAM	84.43	89.01	0.8203	0.8571	0.9130	0.9600
	aiBiG-SAM	84.43	89.01	0.8203	0.8571	0.9130	0.9600
Hypertension	MIVIA	88.18	89.03	0.8332	0.8414	0.9306	0.9368
	iBiG-SAM	88.00	88.70	0.8311	0.8382	0.9291	0.9332
	aiBiG-SAM	88.00	88.70	0.8311	0.8382	0.9291	0.9332

The results clearly demonstrate that **MIVIA** consistently outperforms both iBiG-SAM and aiBiG-SAM across all evaluation metrics.

On the **Heart Disease** dataset, MIVIA achieves the highest test accuracy (90.10%), precision (0.8596), and recall (0.9800), indicating excellent predictive performance and a strong ability to detect positive cases with minimal false negatives.

Similarly, on the **Hypertension** dataset, MIVIA again attains superior results, achieving the highest accuracy (89.03%), precision (0.8414), and recall (0.9368).

These results highlight the robustness and effectiveness of the proposed algorithm. In particular, the improvement in recall demonstrates that MIVIA is especially suitable for

medical classification tasks, where correctly identifying positive cases is of critical importance.

Overall, the experimental findings confirm that the incorporation of inertial and viscosity mechanisms significantly enhances the performance of bilevel optimization algorithms in practical applications.

5. Conclusions

From the experiments conducted, it is evident that MIVIA consistently achieves higher performance in terms of both accuracy and precision compared to iBiG-SAM and aiBiG-SAM. These improvements, although numerically moderate, are meaningful in high-stakes classification tasks such as medical diagnosis, where reducing false positives is crucial.

Furthermore, the experimental results validate the theoretical convergence properties of MIVIA and demonstrate its practical effectiveness in data-driven optimization scenarios. The findings support the conclusion that MIVIA is not only theoretically sound but also empirically advantageous, especially in applications involving noisy, real-world data.

Future experiments may explore the impact of varying the number of hidden neurons, examining convergence speed in terms of iteration count, and extending the analysis to additional datasets or multitask classification problems.

Acknowledgments

This work was partially supported by Chiang Mai University and Fundamental Fund 2026, Chiang Mai University, Chiang Mai, Thailand. The first author would like to thank CMU Presidential Scholarship for the financial support.

References

[1] Parikh, N.; Boyd, S. Proximal algorithms. *Found. Trends Optim.* **1** (2014), 127–239.
[2] Mann, W. R. Mean value methods in iteration. *Proc. Amer. Math. Soc.* **4** (1953), 506–510.
[3] Halpern, B. Fixed points of nonexpanding maps. *Bull. Amer. Math. Soc.* **73** (1967), 957–961.
[4] Moudafi, A. Viscosity approximation methods for fixed point problems. *J. Math. Anal. Appl.* **241** (2000), 46–55.
[5] Ishikawa, S. Fixed points by a new iteration method. *Proc. Amer. Math. Soc.* **44** (1974), 147–150.
[6] Phuengrattana, W.; Suantai, S. On the rate of convergence of Mann, Ishikawa, Noor and SP-iterations for continuous functions on an arbitrary interval. *J. Comput. Appl. Math.* **235** (2011), 3006–3014.
[7] Takahashi, W. Viscosity approximation methods for a countable family of nonexpansive mappings in Banach spaces. *Nonlinear Anal.* **70** (2009), 719–734.
[8] Polyak, B. T. Some methods of speeding up the convergence of iteration methods. *USSR Comput. Math. Math. Phys.* **4** (1964), no. 5, 1–17.
[9] Beck, A.; Teboulle, M. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sci.* **2** (2009), no. 1, 183–202.
[10] Sabach, S.; Shtern, S. A first-order method for solving convex bilevel optimization problems. *SIAM J. Optim.* (2017), in press.
[11] Shehu, Y.; Vuong, P. T.; Zemkoho, A. An inertial extrapolation method for convex simple bilevel optimization. *Optim. Methods Softw.* **36** (2021), 1–19.
[12] Duan, P.; Zhang, Y. Alternated and multi-step inertial approximation methods for solving convex bilevel optimization problems. *Optimization* **71** (2022), 1–30.
[13] Takahashi, W. *Introduction to Nonlinear and Convex Analysis*. Yokohama Publishers, Yokohama, 2009.
[14] Aoyama, K.; Kimura, Y. Strong convergence theorems for strongly nonexpansive sequences. *Appl. Math. Comput*. **217** (2011), 7537–7545.
[15] Goebel, K.; Kirk, W. A. *Topics in Metric Fixed Point Theory*. Cambridge Univ. Press, Cambridge, 1990.
[16] Agarwal, R. P.; O'Regan, D.; Sahu, D. R. Iterative construction of fixed points of nearly asymptotically nonexpansive mappings. *J. Nonlinear Convex Anal.* **8** (2007), no. 1, 61–79.
[17] Aoyama, K.; Kimura, Y.; Takahashi, W.; Toyoda, M. Approximation of common fixed points of a countable family of nonexpansive mappings in a Banach space. *Nonlinear Anal.* **67** (2007), 2350–2360.

[18] Nakajo, K.; Shimoji, K.; Takahashi, W. Strong convergence to common fixed points of families of nonexpansive mappings in Banach spaces. *J. Nonlinear Convex Anal.* **8** (2007), 11–34.
[19] Aoyama, K.; Kohsaka, F.; Takahashi, W. Strong convergence theorems by shrinking and hybrid projection methods for relatively nonexpansive mappings in Banach spaces. *Nonlinear Analysis and Convex Analysis*, 7–26, Yokohama Publishers, 2009.
[20] Bussaban, L.; Suantai, S.; Kaewkhao, A. A parallel inertial S-iteration forward–backward algorithm for regression and classification problems. *Carpathian J. Math.* **36** (2020), 35–44.
[21] Takahashi, W. *Nonlinear Functional Analysis*. Yokohama Publishers, Yokohama, 2000.
[22] Saejung, S.; Yotkaew, P. Approximation of zeros of inverse strongly monotone operators in Banach spaces. *Nonlinear Anal.* **75** (2012), 724–750.
[23] Nesterov, Y. A method for solving the convex programming problem with convergence rate $O(1/k^2)$. *Dokl. Akad. Nauk SSSR* **269** (1983), 543–547.
[24] Huang, G.-B.; Zhu, Q.-Y.; Siew, C.-K. Extreme learning machine: Theory and applications. *Neurocomputing* **70** (2006), 489–501.
[25] Bauschke, H. H.; Combettes, P. L. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. 2nd ed., CMS Books in Mathematics, Springer, Cham, 2017.
[26] Noor, M. A. New approximation schemes for general variational inequalities. *J. Math. Anal. Appl.* 251 (2000), 217–229.
[27] Tibshirani, R. Regression shrinkage and selection via the Lasso. *J. R. Stat. Soc. Ser. B Methodol.* 58 (1996), no. 1, 267–288.
[28] Janosi, A.; Steinbrunn, W.; Pfisterer, M.; Detrano, R. Heart Disease. *UCI Machine Learning Repository*, 1989.

[1] PhD Degree Program in Mathematics, Department of Mathematics, Faculty of Science, Chiang Mai University, under the CMU Presidential Scholarship, Thailand
Email address: serdar_r@cmu.ac.th

[2] Excellence Center in Optimization and Computational Intelligence for Big Data Prediction, Department of Mathematics, Faculty of Science, Chiang Mai University, Chiang Mai, 50200, Thailand
Email address: suthep.s@cmu.ac.th